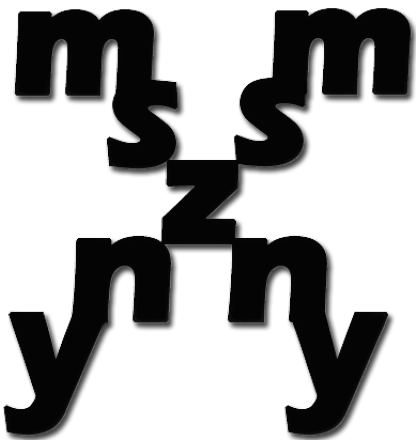
XVI. Magyar Számítógépes Nyelvészeti Konferencia



Szerkesztette:
Berend Gábor
Gosztolya Gábor
Vincze Veronika

Szeged, 2020. január 23–24.

Szerkesztette[1]**:**
Berend Gábor, Gosztolya Gábor, Vincze Veronika
{berendg,ggabor,vinczev}@inf.u-szeged.hu

Felelős kiadó:
Szegedi Tudományegyetem
TTIK, Informatikai Intézet
6720 Szeged, Árpád tér 2.

ISBN: 978-963-306-719-2

Nyomtatta:
JATEPress
6722 Szeged, Petőfi Sándor sugárút 30–34.

Szeged, 2020. január

Az MSZNY 2020 konferencia szervezője:
MTA-SZTE Mesterséges Intelligencia Kutatócsoport

[1] a LaTeX's 'confproc' csomagjára támaszkodva

Előszó

2020. január 23-24-én tizenhatodik alkalommal rendezzük meg Szegeden a Magyar Számítógépes Nyelvészeti Konferenciát. A konferencia fő célkitűzése a kezdetek óta állandó: lehetőséget biztosítani a nyelv- és beszédtechnológia területén végzett kutatások eredményeinek ismertetésére és megvitatására, ezen felül a különféle hallgatói projektek, illetve ipari alkalmazások bemutatására.

Nagy örömet jelent számunkra, hogy a hagyományokat követve a konferencia idén is nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A 39 beküldött cikkből gondos mérlegelést követően 33-at fogadott el a programbizottság, melyek témája számos szakterületet felölel a beszédtechnológiai fejlesztésektől kezdve a legújabb nyelvi elemző eszközök bemutatásán keresztül az orvosi vonatkozású eredményekig.

Az évek során hagyománnyá vált az is, hogy a mesterséges intelligencia vagy a számítógépes nyelvészet egy-egy kiemelkedő alakja plenáris előadást tart a konferencián. Az idei évben Hunyadi László (Debreceni Tudományegyetem) előadásából megtudhatjuk, miként látja a számítógép az emberi viselkedést.

Az idei évben is szeretnénk különdíjjal jutalmazni a konferencia legjobb cikkét, mely a legkiemelkedőbb eredményekkel járul hozzá a magyarországi nyelv-és beszédtechnológiai kutatásokhoz. Továbbá idén második alkalommal kerül kiosztásra a Legjobb Bírálót megillető díj, amellyel a bírálók fáradságos és egyben nélkülözhetetlen munkáját kívánjuk elismerni.

A konferenciához idén is kapcsolódni fog egy kerekasztal-megbeszélés, ahol a főbb szakmai kérdések, a szakterület jelenlegi helyzete és várható haladási iránya, valamint a konferenciához közvetlenül kapcsolódó kérdések kerülnek megvitatásra.

Köszönettel tartozunk az MTA-SZTE Mesterséges Intelligencia Kutatócsoportjának és a Szegedi Tudományegyetem Informatikai Intézetének helyi szervezésben segédkező munkatársainak. Végezetül szeretnénk megköszönni a programbizottság és a szervezőbizottság minden tagjának áldozatos munkáját, ami nélkül nem jöhetett volna létre a konferencia.

A szervezőbizottság nevében,
Ács Judit
Berend Gábor
Novák Attila
Simon Eszter
Sztahó Dávid
Vincze Veronika

Tartalomjegyzék

Szemantika, NLP-eszközök **1**

3 Word Sense Disambiguation for Hungarian using Transformers
Berend Gábor

15 Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben
Kicsi András, Szabó Ledenyi Klaudia, Pusztai Péter, Németh Péter, Vidács László

29 Újabb fejlemények az e-magyar háza táján
Simon Eszter, Indig Balázs, Kalivoda Ágnes, Mittelholcz Iván, Sass Bálint, Vadász Noémi

43 AVOBMAT: a digital toolkit for analysing and visualizing bibliographic data and texts
Péter Róbert, Szántó Zsolt, Seres József, Bilicki Vilmos, Berend Gábor

Beszédtechnológia I. **57**

59 Depresszió detektálása korrelációs struktúrán alkalmazott konvolúciós hálók segítségével
Kiss Gábor, Jenei Attila Zoltán

73 Nagyszótáras beszédfelismerés morfémaalapú rekurrens nyelvi modell használatával
Grósz Tamás

83 A depresszió hang alapú felismerésének optimalizációja hangfelvétel hossz alapján
Azra Pašić, Kiss Gábor, Sztahó Dávid

Poszter, laptopos bemutató **93**

95 FORvoice 120+: magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra
Beke András, Szaszák György, Sztahó Dávid

103 Longitudinális korpusz magyar felnőtt adatközlőkről
Gráczi Tekla Etelka, Huszár Anna, Krepsz Valéria, Száraz Bettina, Damásdi Nóra, Markó Alexandra

115 Szaknyelvi annotációk javításának statisztikai alapú támogatása
Kicsi András, Pusztai Péter, Szabó Endre, Vidács László

129 A tagmondati távolságszámítás módjainak hatása a névmási anaforafeloldásra
Kovács Viktória

141 KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése
Vadász Noémi

155 Automatikus tematikuscímke-ajánló rendszer sajtószövegekhez
Yang Zijian Győző, Novák Attila, Laki László János

Morfológia, helyesírás **169**

171 The Role of Interpretable Patterns in Deep Learning for Morphology
Ács Judit, Kornai András

181 Automatikus ékezetvisszaállítás transzformer modellen alapuló neurális gépi fordítással
Laki László János, Yang Zijian Győző

191 Elírások automatikus detektálása és javítása radiológiai leletek szövegében
Kicsi András, Szabó Ledenyi Klaudia, Németh Péter, Pusztai Péter, Vidács László, Gyimóthy Tibor

205 Szösszenet az elveszett morfémákért - Az alaki analógiák haszna
Naszódi Mátyás

Beszédtechnológia II. **217**

219 Az akusztikus szózsák eljárás korpuszfüggetlenségének vizsgálata
Vetráb Mercedes, Gosztolya Gábor

233 A nyelvkontúrkövető algoritmusok és a gépi tanulás összekapcsolhatóságának vizsgálata
Trencsényi Réka

245 ASR-hibaterjedés vizsgálata a gépi beszédértés szemszögéből
Tündik Máté Ákos, Szaszák György

Poszter, laptopos bemutató II. **259**

261 apPILkáció: egy Android-alkalmazás manysi nyelvtanulás céljára
Bobály Gábor, Horváth Csilla, Vincze Veronika

273 Tárgyas szerkezetek elemzése tenzorfelbontással – áttekintő cikk
Markai Márton

289 A természetesnyelv-feldolgozás fizikai és nyelvi határai
Mészáros Evelin

303 Bu-Bor-éK: grafikus címkenormalizáló eszköz
Novák Attila, Novák Borbála

313 Mély neuronhálós akusztikus modellek súlyinicializálásának vizsgálata
Pintér Ádám, Tóth László, Gosztolya Gábor

323 A történetszerkezet automatikus elemzése és összefüggése az elbeszélő személy érzelmi intelligenciájával
Pólya Tibor

333 Kulcsfogalmak jelentésváltozása a Kádár-korszak politikai diskurzusában
Ring Orsolya, Kmetty Zoltán, Szabó Martina Katalin, Kiss László, Nagy Balázs, Vincze Veronika

343 Automatikus összefoglaló generálás magyar nyelvre BERT modellel
Yang Zijian Győző, Perlaki Attila, Laki László János

Korpusznyelvészet, szintaxis **355**

357 1956 és 1989 között keletkezett propagandaszövegek nyelvi sajátságai
Szabó Martina Katalin, Ring Orsolya, Vincze Veronika

369 Német-magyar nyelvtanulói korpusz (Dulko)
Kappel Péter, Modrián-Horváth Bernadett, Andreas Nolda, Vargáné Drewnowska Ewa

385 Nesze semmi, fogd meg jól! Zéró kopulák automatikus felismerése neurális gépi fordítással
Dömötör Andrea, Yang Zijian Győző, Novák Attila

399 A duplakocka modell és az igei szerkezeteket kinyerő "ugrik és marad" módszer nyelvfüggetlensége, valamint néhány megjegyzés az UD annotáció univerzalitásáról
Sass Bálint

409 Egy emBERT próbáló feladat
Nemeskey Dávid Márk

Szerzői index, névmutató **419**

Szemantika, NLP-eszközök

Word Sense Disambiguation for Hungarian using Transformers

Gábor Berend[1,2]

[1]University of Szeged, Institute of Informatics
[2]MTA-SZTE, Research Group on Artificial Intelligence
berendg@inf.u-szeged.hu

Abstract. In this paper we investigate the applicability of contextual word embeddings for the task of word sense disambiguation (WSD) in Hungarian. We show that a simple k–nn (k–nearest neighbors) approach which relies on multilingual BERT representations can yield highly accurate results in terms of F-scores when evaluated for word sense disambiguation.
Keywords: contextual word representations; multilingual BERT; word sense disambiguation (WSD)

1 Introduction

Word embeddings have been prevalently applied in a variety of natural language processing applications ranging from machine translation (Bahdanau et al., 2014) to information retrieval (Vulić and Moens, 2015) and sentiment analysis (Socher et al., 2013), among others.

A major shortcoming of standard static word embeddings, including `word2vec` (Mikolov et al., 2013) and Glove (Pennington et al., 2014) is that they assign a fixed representation to the individual word forms. That is, the vectorial representations belonging to a word is fixed and it behaves agnostically to the context a particular word is presented. Until recently, such word representations have dominated NLP applications.

Contextualized word representations, such as CoVe (McCann et al., 2017), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), however, have the added favorable property that they are capable of incorporating the context in which a particular word is mentioned upon constructing its vectorial representation. This characteristic of contextualized word embeddings makes them highly appealing for applying them to the task of word sense disambiguation (WSD), where the task is to choose the most appropriate sense a particular word form has based on its context.

There have been some investigation of applying contextual word embeddings for WSD in English (Loureiro and Jorge, 2019; Vial et al., 2019). Our paper is complementary to these results in that here we give a thorough empirical evaluation for using contextual word embeddings for performing WSD in Hungarian.

Our solution uses a simple, yet effective k–nn-based approach for performing WSD. The main contributions of the paper are that

– we evaluate and carefully analyze the applicability of the off-the-shelf multilingual BERT model being applied for Hungarian WSD by a k–nn based approach,
– make the contextualized word embeddings obtained for nearly 12500 sense-annotated utterances publicly available.

2 Related work

One of the key difficulties of natural language understanding is the highly ambiguous nature of language. As a consequence, WSD has long-standing origins in the NLP community Lesk (1986) and it is still in the focus of a series of recent research efforts in NLP (Raganato et al., 2017; Melamud et al., 2016; Loureiro and Jorge, 2019; Vial et al., 2019).

The typical setting for WSD is to categorize the mentions of ambiguous words according to some sense inventory. The most frequently applied sense inventory in the case of English is definitely the Princeton WordNet (Fellbaum, 1998). A Hungarian version of the WordNet also has been created (Miháltz et al., 2008) serving the basis of the Hungarian WSD dataset created by Vincze et al. (2008).

WSD systems either take some unsupervised, knowledge-based or some supervised approach requiring a training corpus with sense-annotated utterances of ambiguous words. Unsupervised approaches could attempt to match the mentions of ambiguous words to their proper sense based on the textual overlap between the context of an ambiguous word and the definitions included to its potential senses according to the sense inventory employed (Lesk, 1986) or be based on random walks over the semantic graph providing the sense inventory (Agirre and Soroa, 2009).

Supervised WSD techniques typically perform better than unsupervised approaches. IMS (Zhong and Ng, 2010) is a classical supervised WSD framework which was created with the intention of easy extensibility. It uses an SVM classifier, which derives features for an ambiguous word based on the word forms and POS tags of the words in its neighborhood. The recent advent of neural text representations have also shaped the landscape of algorithms performing WSD. Melamud et al. (2016) devised the context2vec framework, which relies on a bidirectional LSTM for performing supervised WSD. Most recently, (Loureiro and Jorge, 2019; Vial et al., 2019) have proposed the usage of contextualized word representations for tackling WSD.

Contextualized word representations (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2019) are recent extensions of traditional word embeddings, such as `word2vec` (Mikolov et al., 2013), with the notable distinction that they construct different vectorial representation even for the same word form when employed in a different context. Contextualized word representations employ some language modeling inspired objective and are trained on massive amounts of textual data,

which makes them generally applicable in a variety of settings, including natural language inference (Williams et al., 2018) or reading comprehension (Khashabi et al., 2018).

3 Experiments

We next introduce the dataset we performed our experiments on, as well as the kind of contextual word representations we determined for it.

3.1 The dataset

The dataset we performed our experiments on is derived from the sense-annotated corpus introduced by Vincze et al. (2008). The dataset contains a collection of documents written in Hungarian that are part of the Hungarian National Corpus (HNC) (Váradi, 2002) including mentions towards 39 ambiguous words. The documents are selected from the *Heti Világgazdaság* subcorpus containing mostly news documents related to business and politics. The different word senses got disambiguated in compliance with the sense inventory of the Hungarian WordNet (Miháltz et al., 2008).

The corpus released by Vincze et al. (2008) contains the entire documents in which the sense-annotated ambiguous words are located. The original dataset contains a separate file for each of the word forms in an ISO-8859-1 encoded XML file. We distilled the original WSD corpus (Vincze et al., 2008) into a single and easy-to-handle tab-separated plain text file in UTF-8 format. The distilled version of the dataset differs from the original dataset in that it contains only the local context of the ambiguous words as opposed to the entire document they are included in. We make this dataset accessible [1], a sample line from which is

```
anyagi_a_1_pénzzel_kapcsolatos 1 Az anyagi kár meghaladja az egymilliárd
schillinget .
```

with the first string denoting the ground truth sense label for the ambiguous word, the second item in the line denoting the token position of the ambiguous target word within the excerpt, followed by the excerpt itself in a tokenized format. The entire dataset contains 12477 distinct mentions for one of the 39 ambiguous Hungarian words. The 12477 excerpts contain a total of 449875 tokens.

Figure 1 illustrates the joint distribution of the number of senses per word forms and the Shannon entropy quantifying the heterogeneity of the distributions of the different senses of word forms. We can see that the number of word senses listed for a particular word form ranged between 1 (for the word form *tanár*/teacher) and 14 (for the word form *jár*/go). Perhaps unsurprisingly, a

[1] http://github.com/begab/huWSDdata

strong positive correlation of $\rho = 0.83$ can be observed between the two quantities, i.e. the higher number of distinct meanings a word form has, the higher amount of uncertainty can be observed on average regarding the predictability of its actual meaning in context.

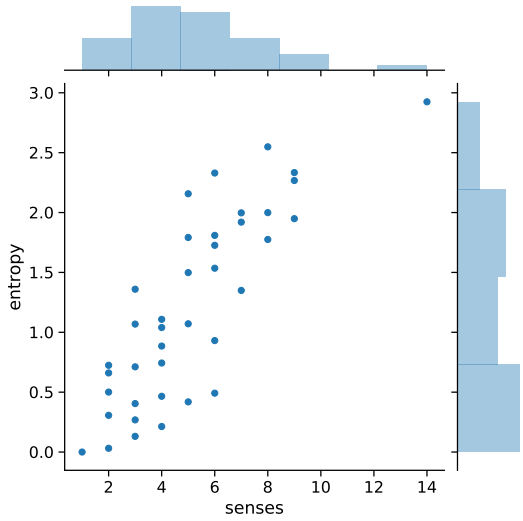


Fig. 1. The joint distribution of the number of distinct senses and the Shannon entropy of their distributions for the 39 word forms in the Hungarian WSD dataset.

3.2 Preprocessing the dataset

We preprocessed the previously introduced WSD dataset using the pretrained cased multilingual BERT (M-BERT) architecture for obtaining contextual word representations. This preprocessing step was conducted using the `Huggingface transformers` Python package (Wolf et al., 2019). We defined the contextualized vectorial form of the individual tokens in the excerpts as the average of the vectorial representations of the word pieces as determined by the M-BERT cased multilingual tokenizer.

The pretrained M-BERT model uses a transformer model which has one word piece-based input layer, followed by 12 stacked layers using self-attention. Each of the 12+1 layers are identical in that they employ vectorial representations of 768 dimensions. We calculated and evaluated the 768-dimensional contextualized word representations for every token. We also performed a sensitivity analysis on using the contextual word representations originating from the different layers of the multi-layered transformer model of M-BERT (cf. Figure 2).

We managed to determine contextual word representations for all but one of the 12477 sense-annotated words in our dataset. The reason why we had to omit one of the sense-annotated words from our analysis was that it was included in an excerpt being longer than the longest sequence M-BERT architecture can possibly deal with, i.e. a sequence length of 512. We also release our contextualized embeddings for the 12476 sense-annotated words that we determined M-BERT representations for at http://github.com/begab/huWSDdata.

3.3 Results

We first review the results obtained in (Vincze et al., 2008) using a traditional approach that is similar to the one applied in IMS (Zhong and Ng, 2010). We subsequently introduce our approach for performing WSD using contextualized M-BERT representations and report our quantitative results.

Overview of the findings from (Vincze et al., 2008) Similar to how it was done in our experiments, Vincze et al. (2008) relied only on the context to be found in the local proximity of the sense disambiguated word forms. The ambiguous words were then represented using the traditional vector space model (VSM) based on the context in the same paragraph of sense-annotated ambiguous words. The features determined for an ambiguous token could additionally include indicator features based on the directly surrounding 3 words of some target word. Vincze et al. (2008) also made use of the POS tag information of the tokens, i.e. they considered only the lemmatized word forms of nouns, verbs, adjectives and adverbs as contextual features from the vicinity of a target token for constructing their feature vector.

Based on the above representation of sense-annotated word forms, Vincze et al. (2008) reports a micro-averaged F-score of 0.703 when relying on a Naïve Bayes classifier and evaluation metrics ranging between 0.727 and 0.749 for applying C4.5 classifier depending on the combination of features they were relying on. Vincze et al. (2008) used a leave-one-out evaluation for assessing the quality of their classifiers for performing WSD. That is, each time a new model was trained on all but one of the feature vectors belonging to the different senses of one of the ambiguous word forms and evaluation was performed against the single one ambiguous instance that was held out from the training instances.

(Vincze et al., 2008) reported evaluation scores for the simple – but often difficult to beat – baseline for always predicting the Most Frequent Sense (MFS) of an ambiguous word, regardless of its context. The MSF baseline obtained an aggregate micro-averaged F-score of 0.694.

Using contextual representations for WSD Our methodology for applying M-BERT representations to WSD is similar to those recently proposed in (Loureiro and Jorge, 2019) for English WSD. An important technical difference between (Loureiro and Jorge, 2019) and our work is that while (Loureiro and Jorge, 2019) based their experiments on the large cased BERT model dedicated

to the English language alone, we were utilizing the multilingual BERT (M-BERT) model in order to be able to use it for WSD in Hungarian. Note that we did not perform any fine-tuning of the M-BERT model to fit the task of WSD, but simply used the pre-trained model in our approach.

The way we evaluated the utilizability of M-BERT embeddings for inclusion in word sense disambiguating the utterances of ambiguous words in Hungarian was via integrating it in a simple k–nn classifier based on the contextualized word vectors determined for the sense-annotated tokens. That is, for a pre-defined value of k and some query word q along with its contextualized word vector $\mathbf{q}$, we simply looked for its k closest neighbor among the sense-annotated contextualized word vectors and returned the majority vote for the sense annotations of the training instances according to their ground truth senses. Similar to (Vincze et al., 2008), we also conducted experiments in a leave-one-out fashion.

We repeated our experiments when relying on different number of nearest neighbors, i.e. $k \in \{1, 3, 5, 7, 9\}$. Figure 2 illustrates the effect of choosing the value for k differently when relying on the M-BERT representation originating from the different layers of the transformer architecture. Figure 2 corroborates previous results on contextual representations that the topmost layers tend to perform better in general, especially for evaluations related to semantics. Results reported in Figure 2 also show a plateauing effect for the last few layers of M-BERT contextualized embeddings. That is, no great improvements can be witnessed when utilizing M-BERT representations derived from the layers in the range of 8 to 12. The earlier layers, however, performed subpar to the final layers.

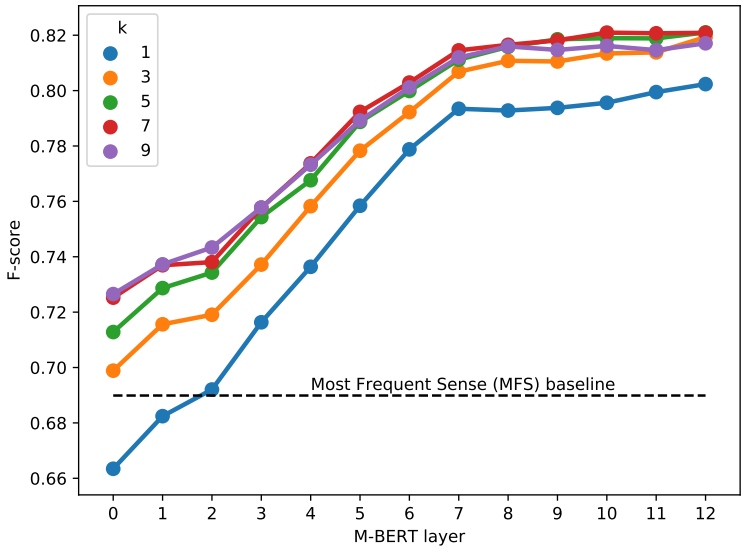
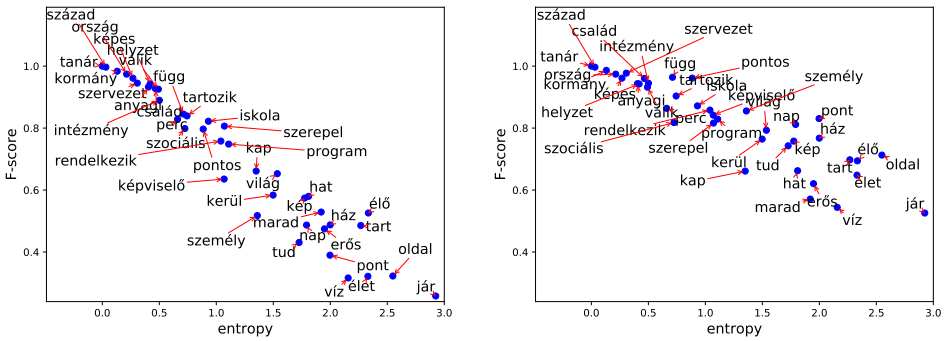


Fig. 2. Aggregated results over the 39 word forms for the MSF baseline and the k–nn model based on M-BERT, when using different values for k.

Figure 2 also shows that increasing the value for the nearest neighbors considered in the prediction can improve performance. Setting k too high, however, is not a good idea, since that would hamper the identifiability of rare senses, and the identification of uncommon senses could often be of potential interest. Hence we argue that using the median from the tested values for k, i.e. $k = 5$, provides a trade-off between delivering increased performance – as opposed to choosing smaller values of k – and being less biased in predicting (the most) frequent senses – as opposed to applying higher values of k.

We can also see it in Figure 2 that k-nn models based on the M-BERT contextual word representations obtained from layer 5 and beyond are outperforming the best reported results in (Vincze et al., 2008) irrespective of the value of k employed. Note that when relying on the final layers of the transformer architecture and employing $k > 3$, we consistently managed to outperform the best previous results by a fair margin (cf. 0.74 versus 0.82).



(a) MFS baseline (b) 5–nn using the last M-BERT layer

Fig. 3. The Shannon entropy of the word sense distributions and the aggregated F-scores of the senses for the individual word forms in the dataset.

As a final assessment, we compared the performance of the MFS baseline and our k–nn solution relying on the M-BERT contextualized representations on the individual level of ambiguous word forms. This comparison contrasts the Shannon entropy of the sense distribution an ambiguous word form has and the F-score obtained for it for a particular model. These results are included in Figure 3 for the MFS and the 5–nn approach relying on the final layer of M-BERT representations for disambiguation.

We can see that while the performance of the MFS baseline fluctuates heavily – with 5 out of 39 word forms having an F-score less than 0.4 – the 5–nn model manages to deliver an F-score at least 0.577, even for the most ambiguous word form (*jár*/go).

We calculated the Person correlation between the results reported in Figure 3. The Shannon-entropy for the sense distribution a word form has and the

performance the different models can achieve for them come hand in hand with a strong negative correlation between the two values. For the MFS and the 5–nn approaches reported in Figure 3 we observed Pearson correlation coefficients of -0.968 and -0.896, respectively. The mere fact that it is more difficult to predict the proper sense for words with a more diverse set of meanings (hence a higher Shannon-entropy) is not so surprising. It would be nonetheless interesting to investigate the reasons for the k–nn based approach behaving less sensitively to the diversity of the ambiguous word forms.

4 Future work and conclusions

This paper focused on WSD in Hungarian for a relatively small set of 39 specific word forms. In order to increase the real world applicability of our model, we plan to extend it to the more challenging all words WSD setting. Training datasets annotated for the all words WSD problem are available in English Raganato et al. (2017); Taghipour and Ng (2015), however, such large scale training data is not currently available for Hungarian at the moment. As a future research, our goal is to investigate how already existing sense-annotated training data – in some possibly foreign language – can improve the performance of WSD.

In this paper, we investigated the extent to which multilingual BERT provides a useful representation for word sense disambiguation. We have seen that a simple solution which uses a k–nn approach for determining the sense of an ambiguous word based on its contextual word representation can obtain highly accurate results.

Acknowledgement

This research was partly funded by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002, supported by the EU and co-funded by the European Social Fund. This work was in part supported by the National Research, Development and Innovation Office of Hungary through the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008).

Bibliography

Agirre, E., Soroa, A.: Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. pp. 33–41. EACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1609067.1609070

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014), http://arxiv.org/abs/1409.0473, cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1423

Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., Roth, D.: Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 252–262. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1023

Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation. pp. 24–26. SIGDOC '86, ACM, New York, NY, USA (1986), http://doi.acm.org/10.1145/318723.318728

Loureiro, D., Jorge, A.: Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5682–5691. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1569

McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: Contextualized word vectors. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 6294–6305. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf

Melamud, O., Goldberger, J., Dagan, I.: context2vec: Learning generic context embedding with bidirectional LSTM. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. pp. 51–61. Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://www.aclweb.org/anthology/K16-1006

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the hungarian wordnet project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)

Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in

Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://www.aclweb.org/anthology/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1202

Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: A unified evaluation framework and empirical comparison. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 99–110. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1010

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), https://www.aclweb.org/anthology/D13-1170

Taghipour, K., Ng, H.T.: One million sense-tagged instances for word sense disambiguation and induction. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning. pp. 338–344. Association for Computational Linguistics, Beijing, China (Jul 2015), https://www.aclweb.org/anthology/K15-1037

Váradi, T.: The Hungarian national corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (May 2002), http://www.lrec-conf.org/proceedings/lrec2002/pdf/217.pdf

Vial, L., Lecouteux, B., Schwab, D.: Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In: Global Wordnet Conference. Wroclaw, Poland (2019), https://hal.archives-ouvertes.fr/hal-02131872

Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian word-sense disambiguated corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008)

Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 363–372. SIGIR '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2766462.2767752

Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1101

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing (2019)

Zhong, Z., Ng, H.T.: It makes sense: A wide-coverage word sense disambiguation system for free text. In: Proceedings of the ACL 2010 System Demonstrations. pp. 78–83. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://www.aclweb.org/anthology/P10-4014

Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben

Kicsi András[1], Szabó Ledenyi Klaudia[1], Pusztai Péter[1,2], Németh Péter[1], Vidács László[1,2]

[1]Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
Szeged, Dugonics tér 13.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
{akicsi,ledenyik,pusztaip,nemethp,lac}@inf.u-szeged.hu

Kivonat Cikkünkben magyar nyelvű radiológiai leletek szövegében automatizáltan azonosítjuk az előforduló testrészeket és elváltozásokat, valamint megállapítjuk a szöveg testrészeinek, elváltozásainak és tulajdonságainak kapcsolatát. Ismertetjük módszereinket, amelyekkel felállítottunk egy megfelelő azonosítóhalmazt, valamint elvégeztük ezek különböző szóalakokhoz való rendelését. Az egyszerű kapcsolatokon kívül az intervallummal vagy utalással megadott speciális eseteket és a tagadásokat is figyelembe vesszük. 487 valós leleten mért eredményeinket közöljük.
Kulcsszavak: radiológia, információkinyerés, NLP, azonosítás, konstituens

1. Motiváció

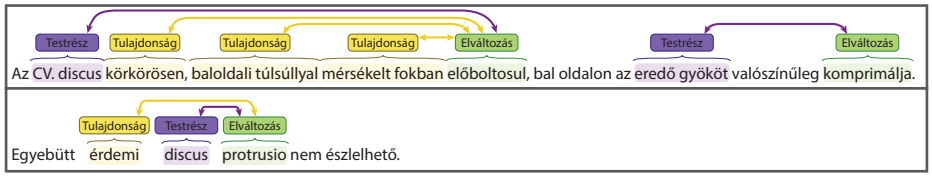
A radiológiai vizsgálatok után az eredmények kezdetben képi formában állnak rendelkezésre. Ezeknek az adatoknak a feldolgozását a mai orvoslásban radiológus végzi orvosi szaktudására támaszkodva. Ez elengedhetetlen a megfelelő értelmezéshez. A radiológus a megtekintett képi információt áttekinti, és szöveges formában rögzíti, általában saját anyanyelvén. Nem csak a képen látott információt írja le, hanem véleményt is alkot, mely a leletben megfogalmazott jelentősebb tények alapján megállapításokat tartalmaz a rögzített elváltozásokról. A leletet és véleményt a radiológus ezután a vizsgálatot kérő szakorvosnak továbbítja, aki ezek figyelembe vételével meghozza a páciens jövőbeli kezelésére irányuló döntéseket. A leletek archiválásra kerülnek, későbbi vizsgálatoknál a radiológus számára elérhetőek, ezzel levonhatóvá válik a következtetés egy korábbi kezelés sikerességéről is, mely szintén kritikus lehet a további lépésekhez.

A leletek tehát rengeteg információt hordoznak, és a radiológus munkájának és szaktudásának gyümölcsét jelképezik. Gépi értelmezésük ezért számos felhasználási lehetőséggel kecsegtet, mint például statisztikák leszűrése, automatikus összehasonlítás a korábbi leletekkel, vélemények automatikus generálása, vagy leletek gyors, vázlatpontos áttekintése. Ezek a jövőben egyúttal tehetnék hatékonyabbá és könnyebbé a radiológus munkáját és szolgálhatnának eszközként a magas szolgáltatási színvonal fenntartása érdekében.

A helyes gépi értelmezéshez azonban feltétlenül szükség van a szöveg elemeinek, entitásainak pontos beazonosítására, tudnunk kell egy testrész leírásáról, hogy pontosan melyik testrészt jelöli, és el kell tudnunk igazodni a megannyi különböző szóalak és szinonima között mind a testrészek, mind a megállapított elváltozások esetében.

2. Áttekintés

Jelen munkában a radiológiai mágneses rezonancia (MR) gerincleletek gépi értelmezésére vonatkozó azonosítási módszereink eredményeit ismertetjük, amelyet a magyarlánc (Zsibrita és mtsai, 2013) nyelvi elemző rendszerrel való feldolgozás alapján alakítottunk ki, és esettanulmányként szolgálhat a hasonló, szakkifejezésekre erősen támaszkodó, szűkös szókincsű természetes nyelvű szövegek gépi értelmezéséhez.



1. ábra: Példamondatok egymásra utalással és viszonylag komplex szerkezettel

A munka során építünk korábbi munkánkra (Kicsi és mtsai, 2019), amely szintén leletek szövegét dolgozza fel. Ebben a szövegben előforduló testrészek, elváltozások és tulajdonságok detektálása volt a célunk. Testrésznek az emberi test egy pontját tekintettük, amely egy viszonylag szűkös, a szöveg alapján alkotóelemekre már nem bonható helyet jelöl. Elváltozásnak tekintettünk minden olyan kifejezést, amely megállapítást fogalmaz meg egy adott testrész állapotáról, illetve annak változásáról. Ide tartoztak a különböző aspektusok is, mint például *„víztartalom”*, amely önmagában nem megállapítás, de a *„víztartalom csökkent”* kifejezés részeként mégis egy elváltozás részét képezi. Szintén ide tartozott a normális állapot megállapítása is, ugyanis a radiológus általában ezen információt is rögzíti, hiszen a károsodás hiánya is értékes információt hordoz egy vizsgálat során. Tulajdonságnak olyan kifejezéseket tekintettünk, amelyek egy elváltozás fokozatát, mértékét vagy egyéb, az elváltozás megnevezéséből nem egyértelmű jellemzőjét írják le, mint például *„3 mm-es”* vagy *„körkörös”*. Ugyanezen nevezéktannal dolgozunk jelen írásban is. A szövegben előforduló tagadásokat a detektálás fázisában nem kezeltük. Detektálási módszerünk gépi tanuláson alapult, melynek során 487 lelet kézzel annotált szövegén tanított Bi-LSTM (Hochreiter és Schmidhuber, 1997) segítségével címkéztük a kifejezéseket, kielégítő eredményekkel.

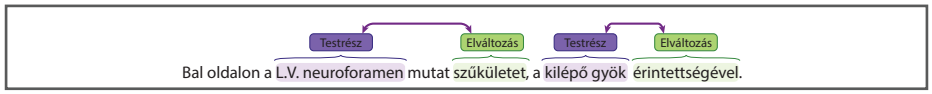
Egy lelet általában állítást fogalmaz meg egy bizonyos testrésszel kapcsolatban. Erre láthatunk két példát az 1. ábrán. Az ábrán jelölésre kerültek a módszerünk által detektált testrészek, elváltozások és tulajdonságok. Azt tehát jól látjuk, és a számítógép számára is egyértelmű már, hogy például a *„CV. discus"* egy testrészt, míg a *„előboltosul"* egy elváltozást jelöl. Az viszont továbbra sem egyértelmű, hogy az emberi test melyik részéről ejtettünk szót, illetve hogy az elváltozásunk pontosan melyik ismert elváltozás, és említése pozitív vagy negatív színezetű-e. Az is megfigyelhető, hogy a tagadás teljes egészében a látókörünkön kívül kerül. A különböző egységek detektálása tehát megtörtént, ám az azonosítás egyáltalán nem, és a mondatok szemantikai jelentése ismeretlen marad.

A fenti problémák tisztán gépi tanulással történő orvoslására nehéz feladat, mely nagy mennyiségű (milliós, vagy milliárdos nagyságrendű), megfelelő minőségű tanítóadat rendelkezésre állása esetén ugyan kivitelezhető lenne, ilyen adatbázisok sajnos még angol nyelvre is nehezen hozzáférhetőek, magyarul pedig még kevésbé. További segítséget nyújthatnának a területspecifikus ontológiák. A jó minőségű angol nyelvű ontológiák nem szabad hozzáférésűek, a szabadon használhatóak pedig egyelőre elmaradnak minőségben a zárt hozzáférésű társaiktól. Ennél is szomorúbb tény, hogy magyar nyelvre tudomásunk szerint átfogó orvosi témájú ontológia nem is létezik. A nagy mennyiségű tanítóadat és a magyar nyelven elérhető ontológiák hiánya miatt az azonosítási és értelmezési feladatainkat nyelvi jellemzők és szabály alapú módszerek alapján végeztük. Jelen írásban ezen megoldásokat tárgyaljuk. Célunk a detektált testrészek és elváltozások azonosítása, kapcsolataik megállapítása, és szemantikai függőségeik feloldása.

3. Módszer

Azonosítási módszerünk egy nyelvi modellen alapul, amelyhez a magyarlánc (Zsibrita és mtsai, 2013) elemző segítségével nyerünk ki bizonyos jellemzőket, majd szabály alapú módszerekkel rendelünk azonosítókat az egyes detektált entitásokhoz. Ide tartozik a szinonimák feloldása is, csakúgy mint az összetartozó latin és magyar szóalakok egymáshoz rendelése. A testrészekhez és elváltozásokhoz egyedi azonosítókat készítettünk, amelyek radiológus által is átnézésre kerültek. Olyan azonosítóhalmazt sem magyar, sem angol nyelvű kapcsolódó kutatásokban, sem nyilvánosan elérhető adatbázisokban nem találtunk, amely elegendő mélységig tartalmazná a gerinc területén található testrészeket és lehetséges elváltozásokat. Az ilyen adatok és ontológiák sajnos még angol nyelvre is kevéssé rendszerezettek, számos kívánnivalót hagynak maguk után az általunk tekintett mélységben. A tulajdonságok azonosításával jelen fázisban nem foglalkozunk, ezek ugyanis általában bonyolultabb szemantikai tartalmat fogalmaznak meg, amely nem feltétlenül írható le előre megalkotott azonosítókkal.

A magyarlánc nyelvi elemző (Zsibrita és mtsai, 2013) a magyar nyelvű szöveg morfológiai, konstituens és dependencia elemzését támogató szoftver, amelyet számos, magyar nyelvű szövegekkel foglalkozó kutatás nagy sikerrel felhasznált már. Az általa biztosított nagyszámú lehetőség közül munkánk során legfőképp a konstituens elemzésre támaszkodtunk, illetve a morfológiai elemzés során feltárt



2. ábra: Egyszerűbb példamondat testrészekkel és elváltozásokkal

tagadásokra. A konstituens elemzés egy fa struktúrát ad, amelyben a mondatok alkotóelemei figyelhetők meg, elkülöníthetők belőle a tagmondatok, amelyek már általában egyetlen szemantikai tartalomra fókuszálnak a nagyobb, összetett mondatok esetében is. Ezt rendkívül hasznosnak találtuk, ugyanis a leletekben szereplő mondatok (például 1. ábra) túlnyomó része egy testrészre mond ki egy elváltozást. A tagadószavak is általában a velük egy tagmondatban lévő entitásokra vonatkoznak, mégpedig pontosan az elváltozásokat leíró szavakra, ahogy a példa második mondatában a *„protrusio”* kerül tagadásra. Ezzel pedig mind a detektált entitások kapcsolata, mind a tagadás tárgya igen könnyen felfedhető. A feladat természetesen nem ennyire egyszerű, rengeteg kivétel felmutatható, ám ezen feltételezések kiindulópontnak mégis több mint elegendők.

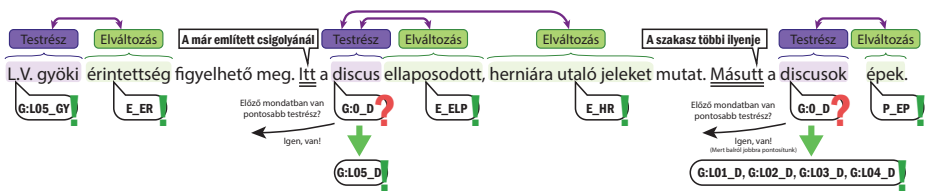
Kézenfekvő megoldás lehetne a dependenciákra támaszkodni a konstituensek helyett, ám a fejlesztés és kísérleteink során azt tapasztaltuk, hogy a konstitunsekre való építés elegendő mélységig tárja fel a szavak egymáshoz tartozását, míg a dependencia a jelenleg kezelt szövegen hibákkal terhelt. Ennek fő oka a speciális nyelvezet lehet, amelyben magyar és latin szavak keverednek. A mondatok azonban itt általában egyszerűbb szerkezettel rendelkeznek egy általános magyar szövegnél, így a konstituensek jó eredményt adnak.

A testrészek és elváltozások azonosítását tehát szabály alapú módszerekkel végezzük. Ehhez felhasználtuk a már meglévő leleteinkben általában szereplő, a detektálás során feltárt testrészeket és elváltozásokat. A szavakat lemmatizáltuk magyarlánc segítségével, ám az eredményeket mindenképpen kézzel kellett javítani, ugyanis itt gyakran speciális szavak fordulnak elő, amelyek a magyarlánc szótárában egyáltalán nem lelhetők fel, és bár az megpróbálja a ragozást így is kimutatni, mégis sokszor problémákba ütközik. Erre lehet példa a *„myelon”* szó, amelyet az elemző egy „myel” szó helyhatározóraggal ellátott változatának tekint. A latin szavak ezen kívül sokszor magyar ragozással fordulnak elő a szövegben (mint például *„herniálódott”* vagy *„degeneratiora”*). Az azonosítók megalkotásánál ezeket először előfordulási szám szerint rendeztük sorrendbe, majd azokat a szavakat tekintettük át, amelyek legalább 10-szer előfordultak a 7648 leletben. Ezek közül lexikografikus listázást követően, kézzel szűrtük ki a helytelenül leírt kifejezéseket. Az összegyűjtött szavak kézzel kerültek csoportosításra, a szóhoz megítélt szótő alapján. Az előálló halmazokhoz azonosítókat rendeltünk. Ezen szóhalmazok közül radiológus segítségével jónéhányat egyesítettünk szinonimák alapján, így például a *„sérv”* és *„hernia”* egy halmazba kerültek. Tapasztalataink szerint az elváltozások azonosításához több orvosi szaktudás volt szükséges, míg a testrészek nehézsége, hogy több, hasonló alakban jelenhetnek meg. Később a listát radiológussal való egyeztetés alapján még kézzel bővítettük.

3.1. Testrészek

A testrészek különlegessége, hogy kapcsolatban állhatnak egymással, amely jelentősen kihat értelmezésükre. A mondat leírhat egy porckorongot *„L. V. discus”*-ként, de mondhatja azt is, hogy *„Az L. V. szerkezete ép. A porckorong apróbb előboltosulása látszik.”*. Ezért nem elég pusztán egymás melletti tokenek sorozatának tekintenünk őket. Utóbbi szerkezetre a megoldás az, ha külön tudjuk detektálni a testrészt, jelen esetben porckorongot, és külön a helyét pontosító másik testrészemlítést, itt az L.V. csigolyát. A testrészeket két részre bontottuk, a csigolyákhoz nem rendelhető és a csigolyákhoz rendelhető testrészekre, utóbbiak pontos helye keresendő. Amikor egy pontosítandó testrészt találunk, akkor egy általánosabb azonosítót rendelünk hozzá, mint például G:0_D, míg ha egy pontos testrészt találunk, akkor egy informatívabb azonosítót kaphat, mint például G:L05. Az általánosabb azonosítóval ellátott testrészeket utólag próbáljuk meg pontosítani. Itt problémát jelent a koreferencia, ahol az *„egyebütt”*, *„máshol”*, *„itt”*, *„többi”* típusú szavak utalnak a testrészre, ahogy az az 1. ábrán is látható. Az itt említett elváltozásokat így egy előző mondatbéli testrészhez kellene vonatkoztatni. Az ilyen szavak detektálásra kerülnek. Az utalások feloldását a 3. ábra szemlélteti. Az ábrán a zöld felkiáltójeles szövegdobozok azt jelzik, hogy a kifejezés megkapta a hozzá illeszkedő, pontos és kellően részletes azonosítót. A piros kérdőjel azt jelenti, hogy csak egy általánosabb azonosítót kapott, ezt próbáljuk feloldani. Detektáltuk az *„itt”* és *„másutt”* szavakat, amelyekhez előre definiált jelentés tartozik. Az „itt” szó egy korábbi testrész csigolyáját, vagy legalább szakaszát jelöli, míg a „másutt” szót úgy tekintettük, hogy egy korábbi csigolya szakaszában a megnevezett csigolyákon kívüli magasságokat jelöli. Ha ezután találunk megfelelően pontosított testrészt az előző tagmondatban, vagy esetleg az előző mondatban, akkor ezek alapján pontosíthatjuk a bizonytalan testrészt. Mivel balról jobbra oldjuk fel az ilyen utalásokat, a példa mindkét dilemmás esetét helyesen fel tudjuk oldani.

További problémát jelenthetnek az intervallummal megadott testrészek, mint például *„L.II.-L. V. discus”*, ahol az intervallum összes eleméről beszél a lelet. Itt a kötőjeleket, az *„és”* és a *„közti”* szavakat keressük, és előfordulásuk esetén átértelmezzük az érintett testrészeket. Ez viszonylag jól automatizálható, ám figyelni kell olyan esetekre is, mint például *„Th.XII.-L.II.”*, ahol a gerincszakaszok közötti váltást is be kellett építeni szabályként.



3. ábra: Példa koreferencia feloldására

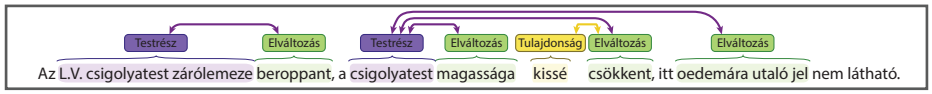
3.2. Elváltozások

Az elváltozások esetében nem igazán fontos két elváltozás kapcsolatát meghatározni. Mivel az aspektusokat a detektálásnál egyben jelöltük az elváltozással, hogy annak valóban értelme is legyen, mint például *„víztartalma csökkent"* esetén, ezért ez az akadály itt nem olyan jelentős. Nagyobb problémát okoz azonban annak értelmezése, hogy pozitív vagy negatív-e az említett elváltozás, tehát az orvos csak megjegyezte, hogy normális állapotot lát, vagy egy valódi degeneratív elváltozást figyelt meg. Ezen megkülönböztetés szintén kézzel került definiálásra. Ezt leginkább a leletek szűkös szókészletének köszönhetően sikerült megfelelő minőségben megtenni. Ez az elváltozások azonosítójában is megjelenik, külön jelöljük a degeneratív elváltozásokat (mint például *„hernia"* - *E_HR*), pozitív állításokat (*„normális"* - *P_NORM*) és az önmagukban polaritással nem rendelkező aspektusokat (*„magassága"* - *ASP_MGS*). Ezeket ismert alakjaiknak megfelelően és magyarlánc segítségével végzett lemmatizálással keressük ki. Mivel a szinonimák már rendelkezésünkre állnak, így ezek tetszőleges szövegben feloldásra kerülnek.

3.3. Kapcsolatok

Bár korábbi elképzeléseink arra irányultak, hogy az entitások közötti kapcsolatokat esetleg gépi tanulási módszerrel állapítanánk meg, úgy találtuk, hogy ezek kézi annotációjára a jelenlegi keretek között nincs feltétlenül szükség. A kapcsolatokat ehelyett a tagmondatokra alapoztuk. Kétféle kapcsolatot kerestünk, testrész és elváltozás, valamint elváltozás és tulajdonság közötti relációkat. A szövegben természetesen előfordulhatnak jelzők a testrészekre is, de ezek az esetek túlnyomó többségében valójában nem is tulajdonságok, hanem elváltozások, mint például *„az előboltosuló discus"* esetében. További megszorító feltételezés, hogy az elváltozások általában egy testrészre, vagy egy testrészek által megadott intervallumra vonatkoznak. Ez szintén helytálló a leletek nagy többségénél, és hasznunkra válik, hiszen így egy elváltozáshoz csak egyetlen testrész keresésére van szükségünk, amelyet a koreferenciák feloldásához nagyon hasonló, priorizált szabály alapú módszerrel valósítottunk meg. A szabályok azonban figyelnek arra, hogy ha *„és"*, *„vagy"* és hasonló szavak választanak el testrészeket, ott minden tagra vonatkozzon az elváltozás.

A leletekben szereplő mondatok tipikusan úgy épülnek fel, hogy először egy testrészt említenek, majd megnevezik a testrész elváltozását, az elváltozás előtt vagy után pedig felsorolják annak tulajdonságait. Ezt megfigyelhetjük például az 1. és 2. ábrán. A mondatok állítmánya gyakran egyik címkéhez sem illeszkedik, mint például *„látszik"* vagy *„észlelhető"*. Természetesen kivételek ez alól a szokás alól gyakran adódnak, ám ezen egyszerű mondatoknál nem nehéz belátni, hogy a kapcsolatok feltérképezése nem komplex feladat. Módszerünk jelen cikk összes példájában szereplő összes kapcsolatot megtalálja. Problémák valójában csak egzotikus megfogalmazás esetén valószínűek, ekkor a kapcsolatot nem sikerül detektálnunk (például *„Mindkét említett discus előboltosul"*). A kapcsolatok hibái általában az entitás detektálás hiányosságaiból erednek.



4. ábra: Példamondat tagadással és több kapcsolattal

3.4. Tagadás

A leletekben gyakran előfordulnak tagadó mondatok is, amik sokszor egy degeneratív elváltozás hiányát írják le. Erre az 1. és 4. ábrán láthatóak példák. Az ismertebb tagadószavakat a magyarlánc is felismeri. Ezek pontos tárgyát is gyakran megadja a dependencia, ezen specifikus szövegeknél azonban azt tapasztaltuk, hogy sokkal jobb eredményeket kapunk, ha ebben is a konstituensekre hagyatkozunk. Ezért itt az előzőekben leírt módszer egy nagyon egyszerű változatát alkalmaztuk, a tagadást tagmondatonként értelmeztük, és amennyiben egy tagmondatban szerepel tagadószó, akkor az egész tagmondatot tagadónak tekintettük. A tagadószó detektálásra a magyarlánc morfológiai elemzését használtuk, ezt azonban ki kellett egészítenünk a *„nincs”*, *„nincsenek”*, *„sincs”* és *„sincsenek”* szavakkal. Egy tagmondat tagadása általában a benne szereplő egyetlen elváltozás jelenlétének tagadása, amely a lelet értelmezése szempontjából kritikus fontosságú. Tapasztalataink alapján így megfelelő eredmények érhetők el a tagadás detektálásában jelenlegi szűk területünk tekintetében.

4. Eredmények

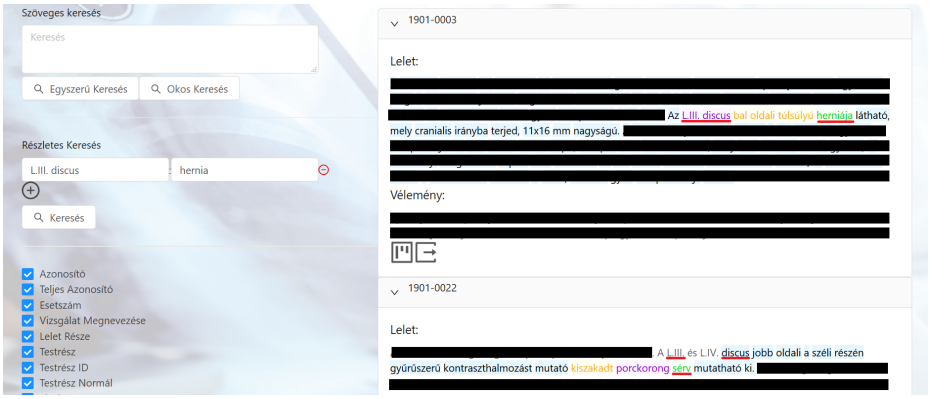
Szabályaink és azonosítóink megalkotása során 5649 lelet adataival dolgozunk, az eredményeinknél közölt számokat pedig a detektálás tanításához is használt 487 leleten végeztük. Módszerünk meghatározott szabályok alapján törekszik testrészek és elváltozások azonosítására és szemantikai kapcsolataik detektálására. Ehhez először is azonosítók szükségesek, amelyeket 5649 lelet adatai alapján alkottunk meg. 39 különböző testrészt különböztetünk meg a csigolyák számait nem tekintve. Ebből 20 testrésznek lehet csigolyaszámozása, tehát ezek mindegyikéhez az általános formán kívül tartozik még 29 pontosítás, ez összesen 629 testrész azonosító. Az elváltozásoknál 214 kóros elváltozást, 8 pozitív jelentésű elváltozást és 20 aspektust különítettünk el, ez összesen 242 elváltozás azonosító.

A 487 leletben 6359 testrész és 7785 elváltozás címke volt. A **testrész** azonosítás során 10258 testrész azonosítót osztottunk ki (ez több, mint az összes detektált testrész, leginkább az intervallumok és a koreferenciák miatt van), ebből 355 különböző. 488 testrészhez nem tudtunk azonosítót rendelni. Az **elváltozás** azonosítás során 9171 elváltozás azonosítót osztottunk ki (itt a többlet az aspektusokból ered, amelyek részei az elváltozásnak), ebből 177 különböző. 332 elváltozáshoz nem tudtunk azonosítót rendelni. Az azonosíthatatlan testrészek és elváltozások természetesen hibát képeznek, ezek nagy valószínűséggel kevésbé gyakori elemek vagy megfogalmazások, ezek javítását a jövőben szintén

kézzel kell megtenni. Szintén ide tartozik a viszonylag nagy mennyiségű, elírások által rontott szóalak, ezek automatikus javítása is utat engedhet a további helyes azonosításhoz. Elmondható azonban, hogy jelenleg a detektált adatokból a testrészek 92,3%-át és az elváltozások 95,7%-át azonosítani tudtuk.

A **koreferencia feloldásban** a vizsgált 487 leletben a következő utalószavak fordultak elő: máshol(376), ebben a magasságban(254), többi(138), itt (122), egyebütt(38), ezen magasságban(34), vizsgált magasságban(20), ugyanebben a magasságban(14), másutt(5), és ugyanitt(3).

A vizsgált 487 leletben összesen 10306 **kapcsolatot** tártunk fel, ebből 6924 elváltozás és testrész, míg 3382 elváltozás és tulajdonság kapcsolata. 843 testrész és 129 tulajdonság volt, amelyhez nem tudott módszerünk elváltozást rendelni. Az elváltozások nélküli testrészek szinte mindegyike abból ered, hogy az egyik testrészt egy másik pontosította, mint például *„L.V. magasságban a discus”*, ilyenkor csak a pontosabb testrészhez kötöttük az elváltozást. A szabadon maradt tulajdonságok nagyrészt a detektálás hibáiból, vagy furcsa megfogalmazásból erednek. 1131 elváltozáshoz nem volt megadva, vagy nem sikerült detektálni egy testrészt sem, itt gyakran mélyebb szemantikai értelmezés lenne szükséges, illetve jó néhány esetben még olvasva sem egyértelmű, hogy milyen testrészhez kötődik egy adott elváltozás. Olyan elváltozások is léteznek, amelyek önmagukban már az érintett testrészre is utalnak. 774 elváltozáshoz nem találtunk egy tulajdonságot sem.



5. ábra: A keresőfelület képernyőképe valós leletekkel. A teljes leletet titkosítottuk, ám a keresés szempontjából lényeges mondatokat meghagytuk

A 487 leletben a magyarlánc konstituens elemzése 6694 tagmondatot tárt fel, ebből módszerünk 1157 tagmondatot tekint **tagadónak**. Nem találtunk olyan valós példát, amelyen a tagadás detektálás hibás eredményt adna, az itt előforduló hibák korábbi feladat hibáiból eredtek minden esetben, mint például az elváltozások detektálásából, vagy a tagmondatokra bontás hiányosságaiból.

Mesterséges példákkal szintén előállíthatók tagadási hibák, külön tagadószavak nélküli megfogalmazásokkal, ám ezeket tapasztalataink szerint nem használják a leletezésben.

Azonosítási módszerünk jelen fázisban már számos felhasználási lehetőséggel bír. Az egyik ilyen lehetőség lehet a leletek intelligens keresése testrészek vagy elváltozások alapján. A módszerre épülő kereső képernyőképe az 5. ábrán látható. A keresőbe beírható keresendő szöveg, ahogy egy hagyományos keresőnél is. Ezen felül azonban testrészek és elváltozások is megadhatók, amelyet kész lehetőségek közül választhatunk, vagy akár sajátot is beírhatunk. Ha a keresődobozra kattintunk, megkapjuk az összes testrész vagy elváltozás listáját, amelyben minden elem csak egyszer (tehát szinonimák nélkül) szerepel. Amennyiben azonban mégis például sérvre szeretnénk keresni hernia helyett, akkor ezt is megtehetjük, ugyanis módszerünkkel ez a keresőszó is kap azonosítót, amely ugyebár megegyezik a hernia azonosítójával. Több testrész és elváltozás is megadható, illetve amennyiben egymás melletti dobozban választjuk őket, a két megadott elem kapcsolatára is szűrünk. Az ábrán jobb oldalon látható két lelet, amelyeken látható, hogy valóban tartalmazzák a keresőszavakat valamilyen formában, és ezek említései kapcsolatban is vannak. Az ábrán látható keresésre egyébként 165 találat volt a 6748 leletből, ezek véletlenszerű sorrendben jelennek meg.

5. Kapcsolódó kutatások

A klinikai szövegek természetesnyelv feldolgozási folyamatában egy fontos lépés, hogy a szavakat kategorizálni tudjuk bizonyos szempontok szerint. Az egyik legalapvetőbb osztályozási forma, amikor a szavakat előre meghatározott névelem osztályokba soroljuk, mint például testrész, betegség, kezelési forma stb. Névelemfelismerés során ugyan meghatározzuk, hogy a mondatban melyik szó milyen osztályba tartozik, ez azonban csak az első lépés a szavak értelmezésének irányába. Az értelmezés vagy azonosítás során az egyik probléma, amivel szembesülhetünk, hogy két ugyanúgy írt szó különböző jelentéssel bír. Ebben az esetben névelemegyértelműsítés segítségével tudjuk feloldani a szavak jelentésbeli különbözőségét. A másik eset, amikor egy jelentéshez, fogalomhoz több különböző szóalak is rendelhető, ilyen esetben névelemnormalizálást alkalmazva, a különböző szóalakokat egy közös fogalomhoz, vagy azonosítóhoz rendelve a probléma feloldható. A nemzetközi szakirodalomban a gyakorlat, hogy a szavakat valamilyen ontológia fogalmaihoz rendelik. Ilyen ontológia például a MeSH (Medical Subject Headings), az RxNorm, a UMLS (Unified Medical Language System) és a SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms), mely a UMLS részét képezi.

A névelemnormalizálás kérdése nagyjából egy idősnek tekinthető a névelemfelismerés problémájával, klinikai szövegeken az első hivatalos megmérettetést a 2013-as ShARe/CLEF eHealth Evaluation Lab Task 1 kihívás jelentette, mely során klinikai szövegekből kellett betegségneveket felismerni és normalizálni (Pradhan és mtsai, 2014). A verseny célja az akkoriban legmodernebbnek számító megvalósítások összegyűjtése és egyben egy alapvonal meghúzása volt ezen a

területen. Az angol nyelvű ontológiák kihasználása érdekében számos eszközt fejlesztettek már, melyek a klinikai szövegekben található releváns kifejezéseket rendelik az ontológiákban található fogalmakhoz. Az ontológiához való hozzárendelést a korai programok javarészt még szabályalapú algoritmusokkal végezték, az elmúlt években azonban folyamatosan jelennek meg az egyre szofisztikáltabb, gépi tanulást alkalmazó modellek. A korai modellek közül néhány említésre méltó példa:

- MedLEE (Friedman, 2000): Szabályalapú eszköz, melyet eredetileg mellkasröntgen leletek feldolgozására fejlesztettek, azóta azonban kiterjesztették a felhasználhatóságát egyéb területekre is.
- MetaMap (Aronson, 2001; Aronson és Lang, 2010): Tudás-intenzív eljárást, azaz természetesnyelv feldolgozási és számítógépes nyelvészeti eljárások egyvelegét alkalmazva rendeli tudományos orvosbiológiai szövegek szavait az UMLS fogalmaihoz.
- cTAKES (Savova és mtsai, 2010): Egy inforációkinyerésre alkalmazható, szabad forrású szoftver, mely többek között használható orvosi szövegekben előforduló kifejezések UMLS fogalmakhoz történő hozzárendelésére is.
- YTEX (Garla és Brandt, 2012): Egy sor kiegészítő modul cTAKES-hez, mely egy általános keretrendszert biztosít szavak ontológiákhoz történő hozzárendeléséhez.
- DNorm (Leaman és mtsai, 2013): Gépi tanulást alkalmazó eszköz, mely hasonlóságot számít a klinikai szövegekben előforduló kifejezések és az ontológia fogalmai között.

Rohit algoritmusa az UMLS-ben található kifejezéseket alapul véve, szerkesztési távolságon alapuló mintázatokat tanult meg, majd ezeket a mintázatokat általánosította korábban nem látott esetek normalizálására (Kate, 2015). Jonnagaddala és szerzőtársai orvosbiológiai szövegekben található betegségnevek felismerésére fejlesztettek CRF (feltételes valószínűségi mezők) alapú névelemfelismerő rendszert, valamint vizsgálták a szótári egyezéskereséses módszerek továbbfejlesztési lehetőségeit, pontosabb névelemnormalizálási eredmények elérése érdekében (Jonnagaddala és mtsai, 2016). Leaman és szerzőtársai a DNorm eszközön alapuló, klinikai szövegekre optimalizált rendszert fejlesztettek, melyet DNorm-C névre kereszteltek (Leaman és mtsai, 2015). A rendszer normalizáláson kívül névelemfelismerést is végez, a klinikai szövegben előforduló kifejezések és az ontológia fogalmai között pedig direkt módon tanul párossával hasonlósági függvényeket. A szerzők állítása szerint a párokban történő tanítás segíti a névelemfelismerő rendszer teljesítményét változatos kifejezéseket tartalmazó szövegek feldolgozásában, valamint a módszer kiterjeszthető más területekre is. A szerzők egy későbbi tanulmányukban elsőként mutatnak be egy semi-Markov modellen alapuló rendszert, mely névelemfelismerést és normalizálást egyidőben végez, mind tanítás, mind pedig kiértékelés közben (Leaman és Lu, 2016). A TaggerOne névre keresztelt rendszer ráadásul szabad forrású. Wang és szerzőtársai saját, kizárólag testrészekből álló ontológiát építettek az UMLS fogalmai alapján, gépi tanuláson alapuló névelemnormalizáló algoritmusuk teljesítményét pedig a Wikipédia tudásbázisára támaszkodó pontozó algoritmussal fejlesztették

tovább (Wang és mtsai, 2019). Az elmúlt évek újabb technológiáit a névelemnormalizálás területén is próbálják alkalmazni, így nem régiben Li és szerzőtársai állítottak fel orvosbiológiai és klinikai szövegek normalizálása terén state-of-the-art eredményeket BERT alapú rendszerükkel (Li és mtsai, 2019).

A magyar nyelvű klinikai szövegeken végzett ide vonatkozó kutatások közül, mindenképp említésre méltó Siklósi és Novák munkája, melyet az orvosi szövegekben található rövidítések megtalálása és feloldása terén végeztek (Siklósi és Novák, 2013; Siklósi és mtsai, 2014; Siklósi és Novák, 2014).

Rendszerünk sajátossága, hogy kevés rendelkezésre álló lelet mellett, valamint magyar nyelvű, területspecifikus ontológia hiányában is képes megfelelő pontosságú névelemazonosítást végezni. Az azonosítás szabályalapon történik, melyhez a leletek szövegét felhasználva egy saját kezdetleges ontológiát is építettünk. Elért eredményeink alapot szolgáltatnak, további kutatások, valamint összetettebb ontológiafejlesztés számára.

6. Összegzés

Cikkünkben azonosítási és információkinyerési feladatokat végeztünk radiológiai leleteken. Korábbi munkánk detektálására is építve azonosítottunk testrészeket és elváltozásokat, amelyekhez saját azonosítóhalmaz definiálására volt szükség. A testrészek, elváltozások és tulajdonságok kapcsolatait is feltártuk, ehhez leginkább konstituens elemzés eredményeire támaszkodva.

Értelmeztük továbbá az intervallumokat, az elváltozások kórosságát, a tagadásokat és utalásokat is. Bemutattuk a módszerrel előállított eredményeinket 487 valós leletre. Munkánknak számos jövőbeli felhasználása lehet a leletek értelmezésében, ilyenek lehetnek az intelligens szemléltetés, strukturált leletek készítése, automatikus véleménygenerálás, vagy, ahogy azt be is mutattuk, intelligens keresés.

Köszönetnyilvánítás

Jelen kutatás az Innovációs és Technológiai Minisztérium ÚNKP-19-3 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült. A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM). Készült az EFOP-3.6.3-VEKOP-16-2017-00002 támogatásával.

Hivatkozások

Aronson, A.: Effective mapping of biomedical text to the umls metathesaurus: The metamap program. Proceedings / AMIA Symposium pp. 17–21 (02 2001)

Aronson, A., Lang, F.M.: An overview of metamap: Historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA 17, 229–36 (05 2010)

Friedman, C.: A broad coverage natural language processing system. AMIA Symposium pp. 270–4 (02 2000)

Garla, V., Brandt, C.: Knowledge-based biomedical word sense disambiguation: An evaluation and application to clinical document classification. Journal of the American Medical Informatics Association : JAMIA 20 (10 2012)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997)

Jonnagaddala, J., Jue, T.R., Chang, N.W., Dai, H.J.: Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion. Database (08 2016)

Kate, R.: Normalizing clinical terms using learned edit distance patterns. Journal of the American Medical Informatics Association : JAMIA 23 (07 2015)

Kicsi, A., Pusztai, P., Szabó Ledenyi, K., Szabó, E., Berend, G., Vincze, V., Vidács, L.: Információkinyerés magyar nyelvű gerinc mr leletekből. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). p. 177–186. Szeged (2019)

Leaman, R., Dogan, R., Lu, Z.: Dnorm: Disease name normalization with pairwise learning to rank. Bioinformatics (Oxford, England) 29 (08 2013)

Leaman, R., Khare, R., Lu, Z.: Challenges in clinical natural language processing for automated disorder normalization. Journal of biomedical informatics 57 (07 2015)

Leaman, R., Lu, Z.: TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models. Bioinformatics 32 (06 2016)

Li, F., Jin, Y., Liu, W., Rawat, B., Cai, P., Yu, H.: Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study. JMIR Medical Informatics 7 (09 2019)

Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., Savova, G.: Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. Journal of the American Medical Informatics Association : JAMIA 22 (08 2014)

Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., Chute, C.: Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA 17, 507–13 (09 2010)

Siklósi, B., Novák, A.: Detection and Expansion of Abbreviations in Hungarian Clinical Notes, Lecture Notes in Artificial Intelligence, vol. 8265, p. 318–328. Springer-Verlag, Heidelberg (2013)

Siklósi, B., Novák, A.: Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – Rövidítések Automatikus Felismerése és Feloldása Magyar Nyelvű Klinikai Szövegekben. In: X. Magyar Számítógépes Nyelvészeti Konferencia. p. 167–176. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2014)

Siklósi, B., Novák, A., Prószéky, G.: Resolving abbreviations in clinical texts without pre-existing structured resources. In: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014). Reykjavík (2014)

Wang, Y., Fan, X., Chen, L., Chang, E., Ananiadou, S., Tsujii, J., Xu, Y.: Mapping anatomical related entities to human body parts based on Wikipedia in discharge summaries. BMC Bioinformatics 20, 430 (08 2019)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. http://rgai.inf.u-szeged.hu/index.php?lang=en&page=magyarlanc (2013)

Újabb fejlemények az e-magyar háza táján

Simon Eszter, Indig Balázs, Kalivoda Ágnes, Mittelholcz Iván, Sass Bálint, Vadász Noémi

MTA Nyelvtudományi Intézet
Budapest, Benczúr u. 33.
{VEZETÉKNÉV.KERESZTNÉV}@nytud.mta.hu

Kivonat A cikkben az e-magyar nyelvfeldolgozó eszközlánc új verziójában, az emtsv-n végrehajtott fejlesztéseket mutatjuk be. Az emtsv fő tulajdonságai közé tartozik a teljes modularitás, amit az egységes formátum és keretrendszer tesz lehetővé. Ebből következik, hogy az emtsv-be könnyen lehet új modulokat integrálni, valamint az egyes elemzési lépéseknél be- és kiszállni. Ezt illusztrálandó egyrészt már létező eszközöket integráltunk (UDPipe, Hunspell), másrészt új modulokat fejlesztettünk (emTerm, emDiff, emZero), harmadrészt a már meglévő modulokat fejlesztettük tovább (detokenizálási funkció az emToken-ben). A cikkben ezeket mutatjuk be, továbbá az emtsv-t teljesítmény és gyorsaság szempontjából összehasonlítjuk hasonló funkcionalitásokkal bíró magyar nyelvfeldolgozó eszközláncokkal, mint a UDPipe, a huspaCy és a Magyarlánc. Az emtsv LGPL 3.0 licenc alatt elérhető a https://github.com/dlt-rilmta/emtsv GitHub repozitóriumból.
Kulcsszavak: e-magyar, emtsv, eszközlánc, erőforrás, tsv, modularitás

1. Bevezetés

Az e-magyar nyelvfeldolgozó rendszer (Váradi és mtsai, 2017) a fejlesztésekor elérhető state-of-the-art (SOTA) eszközöket integrálta egy egységes, könnyen kezelhető, fenntartható és fenntartott eszközláncba. Fontos célja volt, hogy az elkészült eszközlánc a magyar nyelv kutatás- és alkalmazásközpontú felhasználását egyaránt elősegítse, továbbá hogy moduláris és nyílt legyen. Az e-magyar első változatának keretrendszerét a GATE (Cunningham és mtsai, 2011) szolgáltatta, így a modulok közötti kötőszövetet is a GATE belső XML formátuma teremtette meg (Sass és mtsai, 2017). Ennek számos hátránya derült ki a felhasználók visszajelzései nyomán, ezért az e-magyar-nak egy új verziója lett kifejlesztve emtsv néven (Indig és mtsai, 2019b). Az emtsv fontos új tulajdonságait Indig és mtsai (2019a) és Indig és mtsai (2019b) mutatják be – ezek közül itt csak a legfontosabbakat emeljük ki: egységes formátum egy egységes keretrendszerben az xtsv segítségével, teljeskörű modularitás, az elemzőláncba való beszállás és az abból való kiszállás lehetősége a lánc bármely pontján, új modulok egyszerű integrálhatósága, valamint felhőalapú technológiák alkalmazása.

Jelen cikkben az emtsv legújabb fejlesztéseit mutatjuk be. A 2. fejezetben az xtsv-t írjuk le, ami nem csak egy egységes adatformátum, hanem az egész

elemzőlánc keretrendszerét szolgáltatja. A 3. fejezetben az új modulokat, illetve a már meglévő modulokon eszközölt új fejlesztéseket ismertetjük. A 4. fejezetben az `emtsv` könnyebb felhasználhatóságát lehetővé tevő felhőalapú technológiákról szólunk. Az 5. fejezet az `emtsv` egy konkrét projektbeli alkalmazását mutatja be. A 6. fejezetben az `emtsv`-t hasonló szövegfeldolgozó láncokkal vetjük össze teljesítmény és gyorsaság szempontjából. A cikket a 7. fejezetben összegzés és a jövőbeli tervek ismertetése zárja.

2. xtsv

Az `xtsv` keretrendszert az eredetileg az `emtsv`-hez kialakított, de nagyon általános, kiterjeszthető formátum és a formátumot kezelő API együttesen alkotja. Az `emtsv` számára Indig és mtsai (2019b) specifikálták a szükséges API-t és formátumot. Ezt úgy általánosítottuk tovább, hogy az `emtsv` nem nyelvi elemzést végző, általánosítható részeit külön csomagba szerveztük és egységesítettük. Így egy általános keretrendszer jött létre, ami minden, az API-t támogató modullal használható. Ez az `xtsv`, ami elérhető pip csomagként[1] és LGPL 3.0 licenc alatt a GitHubon is[2].

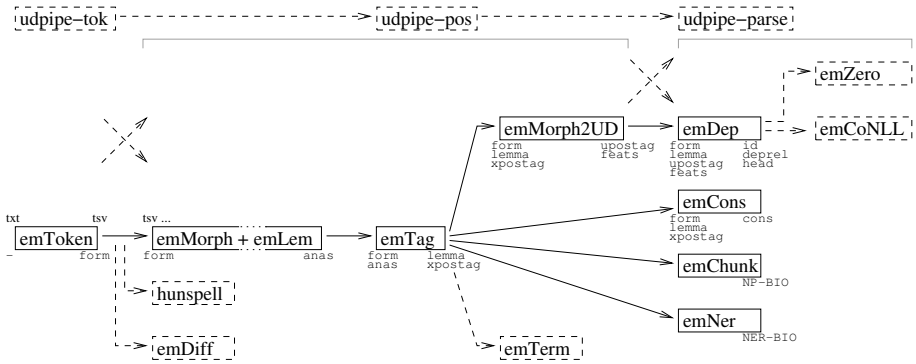
Az `xtsv`-ben használt formátum szembetűnő hasonlóságokat mutat a CoNLL-U Plus formátummal[3], mely a CoNLL-U formátumot[4] kívánja kiterjeszthetővé tenni a kompatibilitás megtartásával. A CoNLL-U Plus viszont egyelőre híján van implementált keretrendszernek.

Az új keretrendszer jellemzői közé tartozik (1) a Javát használó modulok egyszerű konfigurálhatósága, (2) a modulok lusta betöltése (csak azok a modulok töltődnek be, amikre az aktuális futtatás során tényleg szükség van), illetve (3) modulok tetszőleges sorozatának (ún. *preset*eknek) könnyű definálhatósága.

3. Új fejlesztések és modulok

Az `xtsv` lehetővé teszi tetszőleges számú újabb modul beépítését a rendszerbe, amely lehetőséggel éltünk is. Az 1. ábrán láthatjuk a rendszer felépítését; szaggatott vonal jelöli azokat a kiegészítő modulokat, amelyek nem voltak részei az `e-magyar` eszközlánc első verziójának. Az első verzió moduljai: `emToken` tokenizáló, `emMorph` + `emLem` morfológiai elemző és lemmatizáló, `emTag` POS tagger, `emMorph2UD` a morfológiai elemző kimeneti formátumáról a Universal Dependencies (UD) formátumára való átalakítást végző modul, `emDep` dependenciaelemző, `emCons` konstituenselemző, `emChunk` főnévicsoport-felismerő és `emNer` tulajdonnévfelismerő. Ebben a fejezetben az új fejlesztéseket mutatjuk be.

[1] `https://pypi.org/project/xtsv/`
[2] `https://github.com/dlt-rilmta/xtsv`
[3] `https://universaldependencies.org/ext-format.html`
[4] `https://universaldependencies.org/format.html`



1. ábra: Az `emtsv` feldolgozó lánca, a bemeneti és kimeneti mezőkkel. A nyilak értelmezése: minden modul előfeltételezi azt a modulsort, ahonnan hozzá nyíl vezet, de bármely későbbi ponton is lefuttatható.

3.1. UDPipe

A UDPipe (Straka és Straková, 2017) egy olyan eszközlánc, amely CoNLL-U formátumú fájlok elemzését végzi bármilyen nyelvű bemeneti szövegen, amelyen létezik CoNLL-U formátumú tanítóanyag. Az elemzési szintek, amiket megvalósít: tokenizálás, POS taggelés[5] és dependenciaelemzés. A tokenizálásnak természetesen folyó szöveg is lehet a bemenete, de azután az egységes átmeneti formátumot a CoNLL-U adja. Magyarra az egyetlen CoNLL-U formátumú annotált korpusz a Universal Dependencies oldalán található korpusz[6], amely a Szeged Dependency Treebanknek (Vincze és mtsai, 2010) egy kb. 42.000 tokenes része. Ez van train-devel-test alkorpuszokra felosztva, 50-25-25 százalékos arányban. A UDPipe magyar modelljei ezen a tanítókorpuszon lettek tanítva.

Fontosnak tartottuk a széles körben használt szövegfeldolgozó eszközláncok integrálását is az `emtsv`-be, ezért beépítettük a UDPipe-ot, amit annak nyílt forráskódja is támogatott. A UDPipe által megkülönböztetett három szint három önálló modulként jelenik meg az `emtsv`-ben, így ezek bármelyikén be és ki lehet lépni a láncból, és kombinálni is lehet az azonos célú modulokat. (Ezt jelzik az 1. ábrán a keresztező nyilak.) A kényelmes használat kedvéért megoldottuk, hogy a UDPipe által végzett három feladatból kettőt (udpipe-tok-pos, udpipe-pos-parse) vagy az összeset (udpipe-tok-parse) egy lépésben is lehessen futtatni. A kombinálhatóság tekintetében azonban figyelembe kell venni, hogy a magyar nyelvű UD treebank morfológiai címkekészlete a UD 2-es verzióját követi, míg az `emDep` még a UD 1-es verzióját használja. A két verzió között morfológiai

[5] A cikkben egységesen POS taggelésnek hívjuk azt az elemzési szintet, amelynek a kimenete a tokenhez rendelt egyértelmű és teljes morfológiai elemzés (lemma, szófaji címke és inflexiós jegyek).

[6] `https://github.com/UniversalDependencies/UD_Hungarian-Szeged`

szinten alig van különbség[7], de azért számolni kell vele. A két dependenciaelemző kimenete viszont eltér egymástól, mivel az `emDep` a teljes Szeged Dependency Treebanken lett tanítva, ami nem lett átkonvertálva a UD címkekészletére.

3.2. Detokenizálás

Az `e-magyar` első verziójában használt `emToken` tokenizáló eredetileg XML és JSON kimenetet produkált, és ezekben megőrizte a szóköz jellegű karaktereket is. Ezáltal biztosította azt az információt, amivel a kimenet egyértelműen detokenizálható volt.

A `emtsv` egységesen minden modultól az `xtsv` kimenetet várja el, ami egy `tsv`-n alapuló formátum, minden tokent külön sorban ábrázol, és a mondathatárokat üres sorokkal jelzi. Ezért a detokenizálhatóságot ebben az új formátumban másképp kellett implementálni. A kimenet első oszlopa (`form`) tartalmazza a tokeneket, míg a szóköz jellegű szekvenciák megőrzésére új oszlopot vezettünk be (`wsafter`). Amellett döntöttünk, hogy a teljes szekvenciát megőrizzük idézőjelek között, így a kimenet szabad szemmel történő áttekintése során is egyértelmű, hogy hol van és hol nincs szóköz a tokenek után. A szóköz jellegű karaktereket a C nyelvben megszokott escape sorozatokkal ábrázoljuk (`\n`, `\r`, `\t`, `\v` és `\f`).

3.3. Hunspell

A UDPipe-hoz hasonlóan egy másik széles körben ismert eszközt, a Hunspellt[8] is integráltuk az `emtsv`-be. A Hunspell egy nyílt forráskódú helyesírás-ellenőrző, lemmatizáló és morfológiai elemző, amely elsősorban a magyarra és hasonló gazdag morfológiájú nyelvekre lett kifejlesztve. A Hunspell a LibreOffice, OpenOffice.org, a Mozilla Firefox és a Google Chrome beépített helyesírás-ellenőrzője, de zárt programcsomagok is használják. Egy önálló modulként lett az `emtsv`-be integrálva, és – ahogy az 1. ábrán is látható – elkülönül a többi modultól abban az értelemben, hogy a kimenete nem adható át bemenetként egy másik modulnak. Ennek az az oka, hogy a Hunspell egy egyedi morfológiai címkekészletet használ[9], amelyre alkalmazható konverter jelenleg nem elérhető.

3.4. `emDiff`: összehasonlító és kiértékelő modul

A modul célja, hogy két `xtsv` formátumú fájl az `emtsv` egyes moduljai által biztosított elemzési rétegek mentén összehasonlítható legyen. A modul háromféle feladat megoldására alkalmas:

- egyszerű összevetésre,
- az egyik szöveget gold standardként tekintve kiértékelési feladatokra, és
- annotátorok közötti egyetértés számolására.

[7] `https://universaldependencies.org/v2/summary.html`
[8] `http://hunspell.github.io/`
[9] `https://github.com/laszlonemeth/magyarispell/`

Az egyszerű összevetés esetében az `emDiff` csak a szóalakok szintjén végez összevetést, így a tokenizálók kimenetei közötti különbségeket ragadhatjuk meg (pl. az `emToken` és a `udpipe-tok` modul összevetésekor); ehhez a Python `difflib` csomagját[10] használja. A kiértékelési feladatok esetében a két fájl közül az egyiket gold standardnak tekintjük, és ahhoz hasonlítjuk a másik fájl tartalmát. Az `emDiff` a különböző elemzési rétegeket eltérő módszerekkel értékeli ki, így például a tulajdonnév-felismeréshez pontosságot, fedést és F-mértéket számol, míg a függőségi elemzéshez Labeled Attachment Score (LAS) és Unlabeled Attachment Score (UAS) értékeket. Annotátorok közötti egyetértést a címkézési feladatok esetében az `nltk.metrics` csomag[11] által biztosított metrikák (S (Bennett és mtsai, 1954), π (Scott, 1955), κ (Cohen, 1960) és α (Krippendorff, 1980)) alapján számolhatunk. Az `emDiff` minden modul kimenetére használható, vagyis az elemzőláncnak minden ízületére illeszthető (lásd az 1. ábrán).

3.5. `emTerm`: kifejezésannotáló modul

Az `emTerm` tetszőleges forrásból származó, azonosítóval bíró, egy- vagy többszavas kifejezéseket ismer fel és annotál tokenizált, mondatokra bontott és POS taggelt bemeneten. A modulnak kétféle bemenetre van szüksége: az elemzendő szövegre és egy `tsv` fájlra, amely a szövegben jelölendő egy- vagy többszavas kifejezéseket tartalmazza az azonosítójukkal együtt. Egy sorban egy kifejezés szerepelhet, a hozzá tartozó azonosítóval. A többszavas kifejezések szavait @ szimbólum választja el egymástól. A bemenetre egy példa látható az 1. táblázatban.

azonosító	**kifejezés**
`1116804-04,2821`	`etnikai@kisebbség`
`1116832-24,3206`	`képzési@alap`
`47639-52`	`katonai@légi@forgalom`
`47674-1236,28,2821`	`kisebbség`

1. táblázat. Példa az `emTerm` bemenetére. A példában IATE kifejezések és azonosítóik szerepelnek (lásd az 5. fejezetet).

A modul tervezésekor abból indultunk ki, hogy a kifejezések nem nyúlhatnak át mondathatáron, ezért az annotálás mondatnyi egységeken történik. Az annotálást egy egyszerű algoritmus végzi, amely

1. kinyeri a mondatban található összes olyan tokenszekvenciát, amelynek a hossza nem haladja meg a leghosszabb keresendő kifejezés hosszát (pl. a *süt a nap* a következő szekvenciákra tagolódik: *süt*, *süt a*, *süt a nap*, *a*, *a nap*, *nap*);

[10] `https://docs.python.org/3/library/difflib.html`
[11] `https://www.nltk.org/api/nltk.metrics.html`

2. minden tokenszekvenciát olyan formájúra alakít, hogy kereshető legyen a kifejezések között: a szekvencia utolsó tokenjének a szótövét őrzi meg, a többi tokennek pedig a felszíni alakját;
3. ha egy átalakított szekvencia megtalálható a jelölendő kifejezések között, elvégzi a találat annotálását.

A megtalált kifejezések annotációja egy új oszlopba kerül. Az annotáció formátuma `találat_sorszáma:azonosító`. A találatok számozása a mondat elején 1-gyel kezdődik. A többszavas kifejezések esetében az első szónál adjuk meg a teljes annotációt, a kifejezés későbbi szavainál csak a találat sorszámát ismételjük. Ha egy token több kifejezésnek is a része, a találatokat pontosvessző választja el egymástól. Amennyiben az adott token nem része egy keresett kifejezésnek sem, az annotáció cellájába alulvonás kerül. Az annotációs formátumot a 2. táblázat szemlélteti.

ID	token	lemma	szófaj	kifejezés
10	az	az	DET	_
11	etnikai	etnikai	ADJ	`1:1116804-04,2821`
12	kisebbségekre	kisebbség	NOUN	`1;2:47674-1236,28,2821`
13	vonatkozó	vonatkozó	ADJ	_

2. táblázat. Példa az `emTerm` kimenetére. `1`-es sorszámmal szerepel az *etnikai kisebbség* kifejezés, `2`-es sorszámmal pedig a külön azonosítóval rendelkező *kisebbség*. Az azonosítók IATE kifejezések azonosítói (lásd az 5. fejezetet).

Mivel az `emTerm` modul POS taggelt bemenetet igényel, az `emTag` után illeszthető be az elemzőláncba (lásd az 1. ábrán). Ez azt jelenti, hogy a többszavas kifejezések annotálása után bármelyik modul irányába továbbléphetünk újabb nyelvi annotációk hozzáadása céljából.

3.6. `emZero`: zérónévmás-beszúró modul

A modul bemenete a tokenizált, POS taggelt és függőségi elemzéssel ellátott szöveg. Az `emZero` egy szabályalapú rendszer, ami a Szeged Dependency Treebank morfológiai és függőségi annotációján alapul, vagyis az eszközláncnak egy pontjára illeszkedik, az `emDep` mögé (lásd az 1. ábrát).

A program a következő elemeknek a helyére illeszt be zérónévmást:

– finit ige alanyának, ha annak nem volt testes alanya;
– határozott ragozású finit ige tárgyának, ha annak nem volt testes tárgya;
– birtok birtokosának, ha annak nem volt testes birtokosa;
– ragozott és ragozatlan infinitívusz alanyának.

A program egyszerű szabályok mentén végzi az elemek beillesztését, melyek alkalmazása során különböző elemzési rétegek tartalmára támaszkodik (lemma, szófaji címke, függőségi elemzés).

A zérónévmások beillesztése után a mondatfában plusz ágak jelennek meg. A zéró elemek anélkül kapnak saját ID-t, hogy a függőségi elemző által kiosztott ID-ket megváltoztatnák. A kimenetben az alany az ige után, a tárgy az ige és a zéró alany után, a birtokos pedig a birtok után jelenik meg, és egy kombinált ID-t kap, ami az őt megelőző elem ID-jéből és a zéróelem szintaktikai szerepének rövidítéséből (SUBJ, OBJ, POSS) áll. Szófajuk névmás (PRON) lesz, a morfológiai jegyeik között pedig az ige vagy a birtok alapján kiszámolható szám és személy jegyek jelennek meg. A zéró elemek alanyként, tárgyként vagy birtokosként elfoglalják helyüket a mondatfában – erre láthatunk egy példát a 2. ábrán.

```
1       Főként          ADV    3  MODE
2       ötvözőelemként  NOUN   3  OBL
3       használják      VERB   0  ROOT
3.SUBJ  ZERO            PRON   3  SUBJ
3.OBJ   ZERO            PRON   3  OBJ
4       .               PUNCT  0  PUNCT
```

2. ábra: Egy testetlen alany és egy testetlen tárgy beillesztése a függőségi elemzésbe.

4. Felhasználási módok

Komoly elvárás az újonnan megjelenő rendszerek esetében a könnyű futtathatóság és a szolgáltatásként történő könnyű telepítés és skálázódás. Az `emtsv`-ben ezt úgy kívánjuk elérni, hogy az `xtsv`-be belefejlesztettünk egy úgynevezett futtatható Docker módot, amivel az egy paranccsal letölthető Docker image parancssori alkalmazásként használhatóvá válik, ahogy egy `.jar` fájl esetében történne. Ezenfelül lehetőség van a Docker image használatával (vagy nélküle) a REST API-n kereszül elérhető szolgáltatásként elindítani az `emtsv`-t (vagy bármilyen `xtsv`-re alapuló programot), ami a Docker ökoszisztéma elterjedtségének köszönhetően jól skálázható és beépíthető különféle felhős munkamenetekbe.

Szükségesnek tartottuk, hogy a felhasználók a lehető legkevesebb technikai képzettséggel is képesek legyenek használni az `xtsv`-re épülő szolgáltatást. Ezért a keretrendszer része egy főleg demózás és prototipizálás céljára használható webes felhasználói felület, mely segítségével sztenderd HTML interfészen keresztül lehet iterakcióba lépni a szerelőszalaggal. Mivel a háttérben ugyanazt a REST API-t használja a felület, ugyanúgy alkalmas nagy adatok feldolgozására is, mint sok felhasználó által beadott sok kis adat egyidejű kezelésére.

5. Alkalmazás

Az `e-magyar` nyelvelemző rendszert közvetlenül hasznosítjuk az EU által támogatott, hét ország részvételével zajló MARCELL projektben[12]. A projekt célja, hogy elemzett korpuszokat állítson elő a bolgár, a horvát, a lengyel, a magyar, a szlovák, a szlovén és a román nemzeti joganyagból. Terveink szerint ezek a korpuszok *in domain* tanítóadatként fognak szolgálni az EU gépi fordító rendszere számára, melynek fő feladata éppen a különféle jogi szövegek automatikus fordítása.

A jogi szövegek feldolgozása bizonyos esetekben nehezebb a köznapi magyar szövegekhez képest. A feldolgozólánc – elsősorban a függőségi elemző – számára kihívást jelentenek a jogi szövegekben előforduló nagyon hosszú (akár 5.000 szavas!) mondatok. Ezek legtöbbször hosszú felsorolásokat tartalmaznak például kinevezésekről, kitüntetésekről vagy egy másik országgal megkötött vámmegállapodás tételeiről.

A projekt keretében készült el a tokenizáló modul detokenizálási funkciója (lásd a 3.2. fejezetet). Ennek segítségével tudjuk szolgáltatni a megkívánt `SpaceAfter=No` annotációt, mely azt jelzi, hogy az adott token után nem volt szóköz az eredeti szövegben.

Szintén a projekt elvárása volt, hogy megjelöljük a szövegben két jogi terminológiai adatbázis entitásait: az egyik a IATE (InterActive Terminology for Europe)[13], egy többnyelvű terminológiai adatbázis, amelyet az európai intézmények a fordításokhoz használnak; a másik az EUROVOC[14], amely az EU-ban kifejlesztett és rendszeresített, elsősorban jogi fogalmakat tartalmazó tezaurusz. Ezekben az adatbázisokban minden egyes kifejezésnek saját azonosítója van – ezek jelennek meg azonosítóként az `emTerm` bemeneti listájában és persze a kimentében is, ahogy a 2. táblázatban is láthatjuk. A IATE és EUROVOC kifejezések annotálását az `emTerm` modul segítségével, a 3.5. fejezetben leírtak alapján végezzük el.

6. Kitekintés és összehasonlítás

Az `emtsv` szövegfeldolgozó eszközláncnak természetesen léteznek alternatívái, amelyek ki tudják váltani bizonyos funkcionalitásait. Ebben a fejezetben ezeket mutatjuk be és hasonlítjuk össze az `emtsv`-vel az alábbi paraméterek mentén: teljesítmény, gyorsaság, nyelvfüggetlenség, ki- és beszállási lehetőség az egyes lépéseknél, új modulok integrálhatósága és könnyű használhatóság. Ki- és beszállás alatt azt értjük, amikor a felhasználó az elemzőláncot csak egy adott lépésig, például a POS taggelésig használja, az automatikus kimenetet kézzel ellenőrzi, majd a megfelelő formátum betartásával az elemzőláncba vissza tud szállni az immár javított annotációval. Ennek elsősorban az olyan felhasználói

[12] `https://marcell-project.eu/`

[13] `https://iate.europa.eu/home`

[14] `https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc`

forgatókönyvekben lehet jelentősége, ahol fontos a nyelvi annotáció pontossága, hogy elkerüljük a halmozódó hibákat a felsőbb elemzési szinteken.

Az `emtsv`-nek egy alternatívája a már ismertetett UDPipe (lásd a 3.1. fejezetet), ami nyelvfüggetlen és könnyen használható. Az teszi könnyen használhatóvá, hogy csak le kell emelni a polcról a kész tanítóanyagot vagy a már előre tanított modellt. Viszont ugyanez a hátránya is: nagy mértékben függ a tanítóanyagtól. A UDPipe, annak ellenére, hogy a forráskódja szabadon elérhető, nem ad lehetőséget új, saját modulok, például szabályalapú elemzők beépítésére. A modularitás egy másik feltételét, mégpedig azt, hogy az egyes lépéseknél ki- és be lehessen lépni az elemzőláncba, viszont teljesíti, amit az egységes CoNLL-U formátum tesz lehetővé.

Hasonló eszköz a spaCy[15], ami kifejezetten ipari alkalmazásra ajánlott, ezért a könnyű használat és a gyorsaság kritikus fontosságú. Modulárisnak mondható abban az értelemben, hogy támogatja új moduloknak a láncba való integrálását, és az is megoldható, hogy az egyes lépéseknél ki- és be lehessen lépni a láncba. Nyelvfüggő abban az értelemben, hogy ha nincsenek előállítva a megfelelő modellek és erőforrások egy adott nyelvre, akkor arra nem használható. A hivatalos repozitóriumban a magyar még nem szerepel, de számos modell letölthető Orosz György GitHub repozitóriumából[16].

A Magyarlánc (Zsibrita és mtsai, 2013) áll legközelebb az `emtsv`-hez abban az értelemben, hogy mindkettő kifejezetten a magyar nyelvre lett kifejlesztve. A Magyarlánc egy Java-alapú, zártan integrált lánc, amely jelenleg a 3.0 verziójánál tart. A rendszer zártságából fakad, hogy az egyes lépéseknél nem lehet ki- vagy beszállni, és nem lehet új modulokat hozzáadni. Ebből (is) következik, hogy teljesen nyelvfüggő, csak a magyarra működik. Zártságának előnye viszont, hogy könnyen használható és viszonylag gyors. A rendszer a legfrissebb SOTA eszközöket tartalmazza, amelyek részben átfednek az `emtsv` egyes moduljaival.

6.1. Teljesítmény

A teljesítmény kiértékeléséhez a 3. táblázat nyújt segítséget. A morfológiai egyértelműsítés és a lemmatizálás esetében accuracy, a dependenciaelemzés esetében Unlabeled Attachment Score (UAS) és Labeled Attachment Score (LAS), a tulajdonnnév-felismerés (NER) esetében pedig F-mérték a táblázatban megadott mérőszám.

Az utolsó oszlop tartalmazza a SOTA eszközök teljesítményét. Ez az oszlop vonatkozik a Magyarlánc és az `emtsv` teljesítményére is, mivel mindkettőre igaz az, hogy a magyarra elérhető SOTA eszközök lettek beépítve, és ezek az eszközök megegyeznek: a PurePos (Orosz és Novák, 2012) és a dependenciaelemzést végző Bohnet parser (Bohnet és Nivre, 2012) a Szeged Dependency Treebanken tanítva. A morfológiai egyértelműsítés és a lemmatizálás számai a Magyarláncba beépített PurePos teljesítményét mutatják, és magukban foglalják a teljes morfológiai címkézést a lemmatizálással együtt. A dependenciaelemzés számai is a Ma-

[15] `https://spacy.io/`

[16] `https://github.com/oroszgy/spacy-hungarian-models/`

gyarláncba beépített elemző teljesítményét mutatják. A tulajdonnév-felismerés (NER) cellája üres a UDPipe esetében, mert a UD treebankekben nem szerepel tulajdonnévi annotáció. A SOTA szám a NER esetében Simon (2013) eredményein alapul, és a Szeged NER korpusz (Szarvas és mtsai, 2006) 90%-án lett tanítva és 10%-án tesztelve.

A UDPipe fejlesztői kiértékelték az egyes modulok teljesítményét[17] – ezek láthatók az első oszlopban. A UD 2.4-es verziójú treebankjein tanított modellek teljesítményét vettük figyelembe, ezek közül is azokat, amelyeknek nyers szöveg volt a bemenete, mivel a többi elemzőlánc esetében is ez a kezdeti bemenet. A morfológiai egyértelműsítés sorába azt a számot másoltuk át, amely a minden morfológiai címkét tartalmazó (AllTags) elemzés mérőszáma. A UDPipe a UD annotációs sémáját és formátumát követő fájlokkal tud csak dolgozni, ezért a 3. táblázatban látható nem túl magas számok valószínűleg annak köszönhetőek, hogy a magyar nyelvre elérhető egyetlen tanítóanyag nem túl nagy (lásd a 3.1. fejezetet).

feladat	UDPipe	huspaCy	SOTA
morf. egyértelműsítés (ACC)	86,40	94,91	96,33
lemmatizálás (ACC)	88,50	95,49	96,33
dependenciaelemzés (UAS)	72,70	76,18	93,22
dependenciaelemzés (LAS)	67,10	66,58	91,42
NER (F1)	-	93,95	96,10

3. táblázat. Az összehasonlításban részt vevő rendszerek teljesítménye.

A második oszlop a magyar spaCy eredményeit tartalmazza, Orosz György mérései alapján[18]. Azt meg kell jegyeznünk, hogy a morfológiai egyértelműsítésnél szereplő szám itt csak a szófaji címkékre vonatkozik, nem a teljes morfológiai elemzésre. A tulajdonnév-felismerő modul a Szeged NER és a Szeged Criminal NE korpuszok[19] 90%-án, valamint a magyar Hunnerwiki korpuszon (Nemeskey és Simon, 2012) lett tanítva, és a Szeged NER és a Szeged Criminal NE korpuszok 10%-án lett tesztelve. A spaCy weboldalán erőteljesen hangsúlyozzák, hogy a gyorsaság mellett a teljesítmény is megmarad, de a magyar spaCy eredményei – különösen a dependenciaelemzésben – a SOTA eszközök teljesítménye alatt maradnak.

6.2. Gyorsaság

A fenti rendszereket összehasonlítottuk sebesség szempontjából is. Sebesség alatt azt értjük, hogy az adott rendszer egy másodperc alatt hány tokenhez bocsátja

[17] `http://ufal.mff.cuni.cz/udpipe/models`
[18] `https://tinyurl.com/y4ole3ul`
[19] `https://rgai.sed.hu/node/130`

ki a megfelelő szintű nyelvi annotációt. Minden elemzővel kétfajta mérést végeztünk: folyó szöveg feldolgozása POS taggelésig, illetve folyó szöveg feldolgozása dependenciaelemzésig. A méréseket ugyanazon a 100.000 tokenes fájlon végeztük el[20] minden esetben ötször, és az eredményeket átlagoltuk. Az olvasás és írás műveletekhez *RAM disk*-et használtunk, a programok inicializálási idejét figyelmen kívül hagytuk. A mérésekhez a rendszereknek a cikk írásának idejében elérhető legfrissebb verzióját használtuk, és egy erősebb teljesítményű asztali gépen futtattuk.[21] Az eredmények a 4. táblázatban láthatóak.

elemző	POS	dependencia
`emtsv` (CLI)	2.320	300
`emtsv` (REST)	2.600	310
Magyarlánc	5.550	450
UDPipe	9.280	3.300
huspaCy	33.980	15.000

4. táblázat. Az összehasonlításban részt vevő rendszerek sebessége (token/másodperc).

Jól látszik, hogy a huspaCy és a UDPipe mindkét versenyszámban jelentősen gyorsabb, mint a többi program. Sőt, a huspaCy-ról nyugodtan állíthatjuk, hogy nagyságrendekkel gyorsabb, mint a versenytársai. Az `emtsv`-nek a kliensszerver (REST) módban való futtatása nagyobb teljesítményt biztosít, mint a parancssoros (CLI) használata. Az `emtsv` Docker verziója körülbelül 300 token/másodperccel teljesít rosszabbul a táblázatban feltüntetett értéknél a POS taggelésben, és körülbelül 20 token/másodperccel a dependenciaelemzésben.

6.3. Összehasonlítás, diszkusszió

A paraméterek, amelyek mentén az egyes rendszerek összehasonlítását végeztük, az alábbiak: teljesítmény, gyorsaság, nyelvfüggetlenség, ki- és beszállási lehetőség az egyes lépéseknél, új modulok integrálhatósága és könnyű használhatóság. Mindezen paramétereket már részletesen kifejtettük a megelőző fejezetekben, itt egy összefoglaló táblázatban a teljes képet tekintjük át. Az 5. táblázatban nyújtjuk a fentebb leírtak mátrixos ábrázolását.

A könnyű használhatóság mindegyik rendszerre igaznak bizonyult, így annak megkülönböztető ereje nincsen. Az egyes elemzési szinteknél való be- és kilépési lehetőség mindegyik rendszerre áll, kivéve a Magyarláncot, ami teljesen zárt ebből a szempontból. A nyelvfüggetlenséget szigorúan véve igazából egyik rendszer sem teljesíti, de a UDPipe minden olyan nyelvre használható, amire létezik UD treebank, ami jelenleg 90 nyelvet jelent, ami messze felülmúlja az összes általunk

[20] Kivéve a huspaCy-t, ami túl gyors volt ahhoz, hogy megbízható eredményt adjon 100.000 tokenen, ezért itt 1.000.000 tokenen végeztük ezt a mérést.

[21] CPU: 4 mag, 4 GHz; RAM: 16 GB

	teljesítm.	gyors.	nyelvfügg.	ki-be	integrálh.	használh.
`emtsv`	O	X	X	O	O	O
Magyarlánc	O	X	X	X	X	O
UDPipe	X	O	O	O	X	O
huspaCy	X	O	X	O	O	O

5. táblázat. A rendszerek összehasonlító táblázata a vizsgált paraméterek mentén (teljesítm. = teljesítmény, gyors. = gyorsaság, nyelvfügg. = nyelvfüggetlenség, ki-be = ki és belépési lehetőség, integrálh. = új modulok integrálhatósága, használh. = könnyű használhatóság). O-val jelöljük azt, ha a rendszer az adott paraméter megvizsgálásából pozitívan jön ki, X-szel, ha nem.

ismert, magát nyelvfüggetlennek állító rendszer teljesítményét. A teljesítmény és a gyorsaság fordított arányosságban tűnik lenni, vagyis nincs olyan rendszer, ami egyszerre valósítja meg ezt a két kontradiktórius elvárást.

7. Összegzés és jövőbeli tervek

A cikkben az `e-magyar` újabb verzióján, az `emtsv`-n végrehajtott újabb fejlesztéseket mutattuk be. Az újonnan fejlesztett modulok, reményeink szerint, még több jövőbeli kutatáshoz tudnak segítséget nyújtani. A fejlesztések során kifejezetten szem előtt tartottuk a bölcsészet- és társadalomtudományok művelőit is, hogy számukra is praktikus alkalmazásokat hozzunk létre.

A teljes elemzőlánc elérhető LGPL 3.0 licenc alatt a rendszer GitHub repozitóriumában[22]. Mivel az `e-magyar` folyamatos fejlesztés alatt áll, cikkünkben a fejlesztés folyamatának csak egy pillanatképét tudjuk adni. Ebből az következik, hogy a fenti repozitóriumot érdemes újra és újra felkeresni.

Az `xtsv` keretrendszer megalkotásakor szükségünk volt egy olyan egységes formátumra, ami a kötőszövetet tudja alkotni az egyes modulok között, és nem kizárólag belső szabvány, hanem igazodik a nemzetközileg használt, elterjedt formátumokhoz is. Az egyik ilyen a cikkben többször is említett CoNLL-U, aminek több hátránya is mutatkozott, ezért dolgoztuk ki az `xtsv` formátumot. Időközben a Universal Dependencies közösség fejlesztői kidolgozták a CoNLL-U-nak egy kiterjesztett verzióját, a CoNLL-U Plust, ami számos tulajdonságában megegyezik az `xtsv`-vel, így ha minden paraméterében megfelel, lehet, hogy az `e-magyar`-t átállítjuk a CoNLL-U Plus formátum használatára.

Hosszabb távú terveink között szerepel egy olyan nagyméretű tanítókoprusz létrehozása, amely a UD v2-es formátumát és címkekészletét követi. Ezzel képesek lennénk teljesen UD-kompatibilis dependenciaelemzés kibocsátására. Hozzátesszük, hogy ez a teljes magyar nyelvtechnológia számára nagyon fontos előrelépés lenne.

[22] `https://github.com/dlt-rilmta/emtsv`

Köszönetnyilvánítás

A kutatást az Európai Bizottság CEF Telecom programjában nyertes 2017-EU-IA-0136 számú MARCELL nevű projektje támogatta.

Hivatkozások

Bennett, E.M., Alpert, R., Goldstein, A.C.: Communications Through Limited-Response Questioning. Public Opinion Quarterly 18(3), 303–308 (1954)

Bohnet, B., Nivre, J.: A Transition-based System for Joint Part-of-speech Tagging and Labeled Non-projective Dependency Parsing. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1455–1465. EMNLP-CoNLL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)

Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6) (2011)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., Makrai, M.: One format to rule them all – the `emtsv` pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 155–165. Association for Computational Linguistics, Florence, Italy (2019a)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: `emtsv` — egy formátum mind felett. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2019b)

Krippendorff, K.: Content Analysis: An Introduction to its Methodology. Sage, Beverly Hills, CA (1980)

Nemeskey, D.M., Simon, E.: Automatikus korpuszépítés tulajdonnév-felismerés céljára. In: Tanács, A., Vincze, V. (szerk.) IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013). pp. 106–117. Szeged (2012)

Orosz, Gy., Novák, A.: PurePos — an open source morphological disambiguator. In: Sharp, B., Zock, M. (szerk.) Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science. p. 53–63. Wroclaw (2012)

Sass, B., Miháltz, M., Kundráth, P.: Az `e-magyar` rendszer GATE környezetbe integrált magyar szövegfeldolgozó eszközlánca. In: Vincze, V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 79–90 (2017)

Scott, W.A.: Reliability of content analysis: The case of nominal scale coding. The Public Opinion Quarterly 19(3), 321–325 (1955)

Simon, E.: Approaches to Hungarian Named Entity Recognition. Ph.D.-értekezés, PhD School in Cognitive Sciences, Budapest University of Technology and Economics (2013)

Straka, M., Straková, J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (2017)

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: A highly accurate Named Entity corpus for Hungarian. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). pp. 1957–1960. ELRA (2006)

Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: Az `e-magyar` digitális nyelvfeldolgozó rendszer. In: Vincze, V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 49–60 (2017)

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010. ELRA, Valletta, Malta (May 2010)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)

AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts

Róbert Péter[1], Zsolt Szántó[2], József Seres[2],
Vilmos Bilicki[2], Gábor Berend[2,3]

[1] University of Szeged, Institute of English and American Studies
[2] University of Szeged, Institute of Informatics
[3] MTA-SZTE, Research Group on Artificial Intelligence
robert.peter@ieas-szeged.hu
{szantozs,bilickiv,berendg}@inf.u-szeged.hu,
seres.jozsef.1@stud.u-szeged.hu

Abstract: The objective of this paper is to demonstrate the different functions and features of the multilingual AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) data-driven digital tool. This new web application enables digital humanists to critically analyse the bibliographic data and texts of large corpora including entire library repositories and digital collections at scale. The implemented state-of-the-art text analytical and visualization tools such as topic modeling and network analysis provide interactive close and distant reading of texts and bibliographic data. The export functions of AVOBMAT facilitate the reproducibility of the results and transparency of the preprocessing and text analysis.

1 Introduction

Text and data mining methods have become increasingly widespread in the natural, social sciences and the humanities in recent years (Prescott, 2015; Graham et al., 2016; Moreux 2016; Schiuma and Carlucci 2018; Mahmoudi and Abbasalizadeh, 2019).Vast amount of humanities resources has been digitalized and encoded in the last three decades in various quality and made available in numerous open access and subscription-based databases with many restrictions including the lack or limited access to the digital collections for exploratory analysis with data-driven methods and tools. It can be observed that the often rich bibliographic (meta)data of documents are scarcely and partially made use of during the corpus preparation and computational text analysis process. This is not surprising since, as Franco Moretti, Matthew L. Jockers and Mikko Tolonen have argued, bibliographic metadata has been unmapped as a means of investigating (literary) history (Jockers, 2013; Moretti, 2013; Tolonen, 2015). Recent research has demonstrated that, like text mining, the critical examination of bibliographic (big) data can also offer many new insights, unveil hitherto overlooked patterns and trends, provide novel type of evidence and findings as well as question old hypotheses in the humanities (Varlamis and Tsatsaronis, 2011; Péter,

2011, Fenlon et al., 2012; Jockers, 2013; Prescott, 2013; Péter, 2015; Hill et al. 2019; Lahti et al., 2019). That is why the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) research tool combines these two data-driven approaches in one integrated, interactive and user-friendly web application. It hopefully assists non-programmer researchers in their critical text and data mining of open access text corpora, individual digital collections and bibliographic datasets. The implemented state-of-the-art analytical and visualization tools ranging from network analysis to topic modelling can provide close and distant reading of texts and bibliographic data at scale. The export of the preprocessed texts and analysis results facilitates the reproducibility of the findings and foster critical thinking about the text preparation and analytical process.

Since March 2017 researchers, library IT specialists and students in the Institute of Informatics, Klebelsberg Library and Department of English Studies at the University of Szeged have been working on the development of the AVOBMAT research tool. The aim of this paper is to briefly demonstrate the workflow and the different analytical functions and features of the AVOBMAT web-based application designed for digital humanities research.

2 Related work

Text mining and natural language processing techniques have been utilized for the analysis of linguistic corpora and databases for a long time. However, with very few exceptions such as the Media Monitoring of the Past project[1], most publicly available historical, literary and cultural databases do not allow researchers to analyse the included texts with data-driven and natural language processing methods in a critical manner. In commercial products such as the Gale Digital Scholar Lab[2] (inaccessible at most universities and research institutions) individual or library databases, (pretrained) NLP models and dictionaries cannot be uploaded. Furthermore, most currently available web-based text analysis tools such as Voyant Tools[3] cannot cope with large corpora of texts (Sinclair and Rockwell, 2019). The relatively few browser-based digital humanities applications only allow users to set a reduced number of processing parameters for the NLP algorithms they offer. Paper Machine[4] as a plugin for Zotero Standalone bibliographic management software once managed to combine basic bibliographic metadata (date, title and location) and topic modeling analysis but it is not compatible with the current version (5.0) of Zotero and thus it is no longer maintained. The Interactive Text Mining Suite[5] web application pre-processes texts (txt, pdf) and provides cluster, topic and frequency analyses but it can handle few metadata fields such as author, year, title and category (Scrivner and Davis, 2016; Scrivner and Davis, 2017; Scrivner et al., 2017). Among these browser-based applica-

[1] https://impresso-project.ch/app/#/
[2] https://www.gale.com/primary-sources/digital-scholar-lab
[3] https://voyant-tools.org/
[4] http://papermachines.org/
[5] https://languagevariationsuite.wordpress.com/2016/03/18/interactive-text-mining-suite-itms/

tions, Lexos[6] provides the greatest number of tools for pre-processing and segmenting the uploaded texts (TXT, HTML and XML). Lexos tokenizes texts, identifies n-grams, generates statistical summaries, visualize the results (word cloud, multicloud, bubbleviz), analyses the digitized texts (frequency, k-means, clustering, cosine similarity ranking). It also provides a "topic cloud" based on the output format created by MALLET (McCallum, 2002) data but one cannot perform topic modeling within Lexos itself (Kleinman et al, 2019). Unlike these tools, AVOBMAT can enrich, analyse bibliographic data and filter the uploaded digital collections for built-in computational text analysis with the help of metadata or (full-text) keyword searches including fuzzy and proximity queries. The latter filtering can also be used in the AVOBMAT system to explore and visualize the bibliographic data of the filtered digital collections in an interactive and dynamic way.

3 Upload, preprocessing and metadata enrichment

Users can currently upload Zotero collections in CSV and RDF (with full texts) formats as well as EPrints (library) repositories (EP3 XML with urls to the full texts). Zotero can import 20 different types of bibliographic formats (e.g. MARC, BibTex) that users can organize in collections. In Zotero users can modify metadata and texts of these collections, for example, by manually adding new items. These collections can be imported in AVOBMAT in CSV and RDF formats. The Zotero-based csv structure containing 87 bibliographic (meta)data fields (e.g. author, publisher, title) was expanded with 20 other new fields such as bookseller, printers, gender (of the authors) and frequency of publications that researchers can analyse in various ways (see chapter 4) and make use of them when creating their own digital collections to be explored in AVOBMAT.

The differences between the various bibliographic metadata schema standards are reconciled. For example, the EP3 XML 'Publication' field is identical with the Zotero 'Publication Title' field, thus they are both imported into a common Elasticsearch field as publicationTitle. Through the csv upload AVOBMAT can also import full texts of documents via either the added "Full Text" field or the "Url" field pointing to the full texts of the documents. Hence it can import texts in several formats including TXT, PDF, DOC(X) and XML since the Apache Tika Python library converts them to plain text.

In the preprocessing phase the user can set the individual parameters for the different types of content analyses including N-gram viewer, topic modeling, significant text and TagSphere. Users can create different configurations for the different analyses and visualizations. The following preprocessing steps are implemented: (i) lowercase the text; (ii) remove numbers; (iii) remove non-alphabetical tokens; (iv) replace expressions and characters; (v) remove hyphens. The (vi) context-filter can be used to keep the context of a keyword or keywords, and remove all other parts of the document. Users can specify the search terms and the length of the context.

[6] http://lexos.wheatoncollege.edu/upload

The preprocessing interface also provides three optional functions (vii) lemmatization, (viii) punctation and (ix) stopword filtering) that can be configured for each analysis separately. In these cases the outcome of the module depends on the language of the texts to be uploaded. There are two ways to assign a language to a document—researchers can either manually select a language for the full dataset or choose the automatic language detection option. In case of automatic language detection, the system chooses a language independently for each document by using the langdetect language detection library[7]. Based on the selected or detected language one can select stopword, punctuation filtering and lemmatization. For the stopword and the punctuation filtering we use the spaCy library[8]. Extra stopword and punctuation lists can also be added. In case of Bulgarian, Czech, Estonian, Hungarian, Italian, Romanian, Serbian, Slovenian, Polish and Spanish we use the LemmaGen (Juršic et al., 2010) – an open source multilingual tool – for lemmatization, and for other languages with spaCy support – including English, French and German – we use spaCy. All the content analyses are performed online based upon the preprocessed texts. The raw and the preprocessed texts are stored in distinct fields on the server. Hence, one can access and search the raw full texts after the preprocessing phase.

The metadata enrichment includes the automatic identification of the gender of the authors and automatic language detection of the documents. We utilize a further developed version of Python sexmachine package for automatic identification of the gender of the authors primarily based on their forenames.

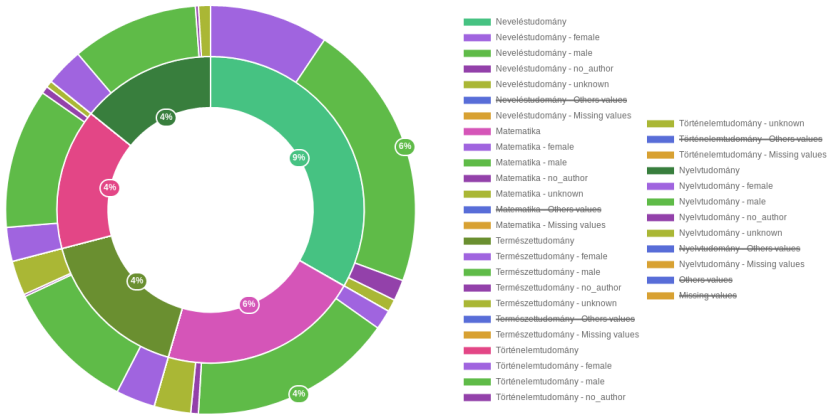


Fig. 1. Distribution of female, male, unknown and no author according to major disciplines as recorded in the Acta repository of the University of Szeged (43300 articles)

[7] https://github.com/shuyo/language-detection
[8] https://spacy.io/

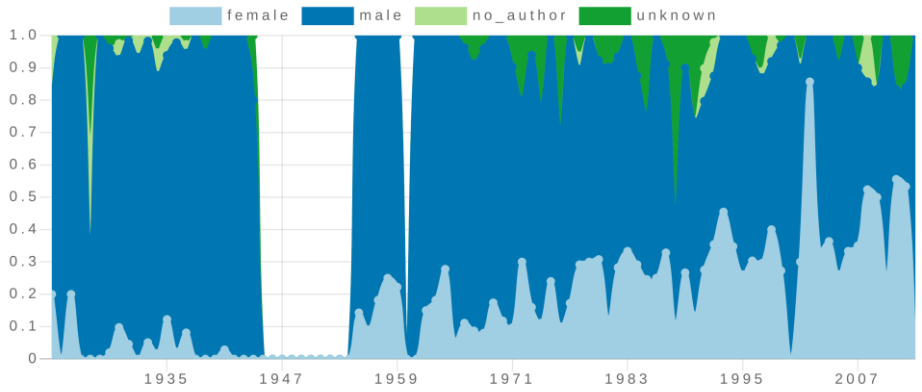


Fig. 2. Distribution of female, male, unknown and no authors in 1653 linguistics journal articles as recorded in the Acta repository of the University of Szeged between 1924 and 2013.

Empowered by the Elasticsearch engine users can search and filter the uploaded and enriched bibliographic data and preprocessed texts in faceted, advanced and command line modes and perform all the subsequent analyses on the filtered dataset. Through the search results interface it is possible to access the uploaded texts and related metadata directly. Our framework also supports fuzzy and proximity search as well as regular expression queries.

The reproducibility and transparency of the experiments and results are enhanced by the export of the parameter settings, preprocessed texts, generated tabular statistical data of the analytical results and visualizations in different formats. The application also enables researchers to use the processed texts and bibliographic data in other software. AVOBMAT does not retain any data after a session has expired.

4 Bibliographic data analysis

Having filtered the uploaded databases and selected the metadata field(s) to be explored, users can, among others, (i) analyse and visualize the bibliographic data chronologically in line and area charts in normalized and aggregated formats; (ii) create an interactive network analysis of maximum three (meta)data fields; (iii) make pie as well as horizontal and vertical bar charts of the bibliographic data of their choice according to the provided parameters: researchers can choose the metadata field(s) and the number of top items for visualization. For example, AVOBMAT can create an interactive network graph of authors, publishers and booksellers. All the analytical results and visualizations such as graphs, time series and charts can be exported in csv and png formats. To foster the critical investigation of the bibliographic records it also presents the number of missing and other values (not included in the dataset limited by the selected number of top items parameter) in the filtered corpus.

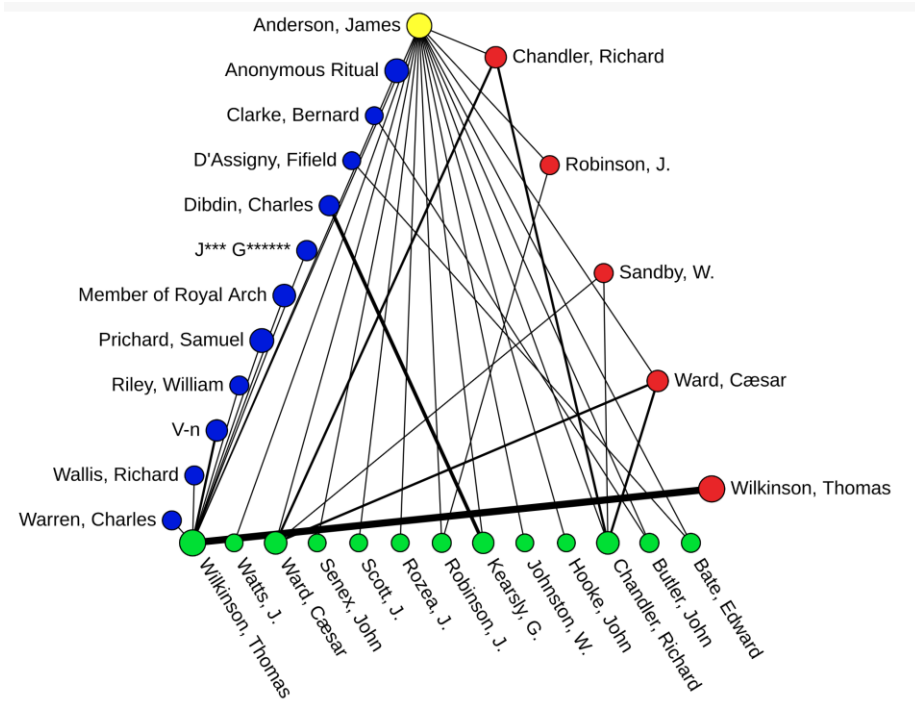


Fig. 3. James Anderson's (author of the *Constitutions* of Freemasonry) eighteenth-century publisher-bookseller network

5 Content analysis

5.1 N-gram viewer

The diachronic analysis of texts is supported by the N-gram viewer in AVOBMAT. It shows the yearly count of the specified n-grams. The inputs are the list of n-grams and a selected metadata field (e.g. the full texts or the titles of the documents). The n-grams with a maximum 5-word length are generated at the preprocessing stage. The chart also provides a normalized view where the number of the n-grams is divided by the total number of words in the given years.

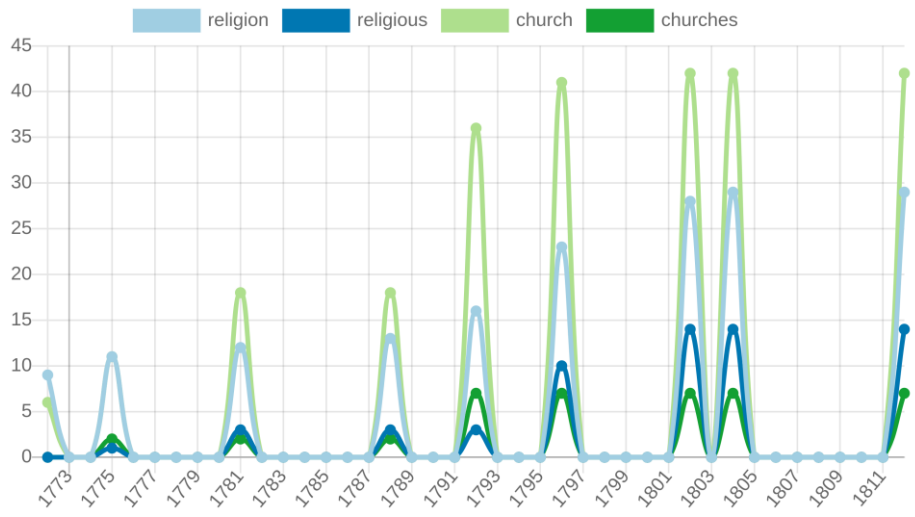


Fig. 4. The terms religion, religious, church and churches in the nine different editions of William Preston's *Illustrations of Masonry*, 1775-1812.

5.2 Word clouds

Word clouds can be efficient tools to highlight prominent terms in a corpus. We implemented two special types of word clouds: the significant text cloud showing what differentiates a subset of the documents from others[9] and the TagSpheres (Jänicke and Scheuermann, 2017) enabling users to investigate the context of a word. There are bar chart versions of the different word clouds that present the applied scores and frequencies.

The significant text visualization highlights the most related terms to a special query. If the user filters a time period or selects an author by using the search functions of the AVOBMAT, this tool highlights the words that are most strongly related to this selected subset of documents.

Traditional word clouds treat the words independently and lose the contextual information between them. For the graphical representation of the context of a word in a corpus the TagSpheres was integrated. It creates tag clouds that show the co-occurring words of a specified search term in a corresponding word distance. Besides the search term, the user can specify the minimum frequency and the maximum distance of the co-occurring words. While the minimum frequency handles the rare, uncommon words and maximum word distance controls mostly the size of the chart.

[9] https://www.elastic.co/guide/en/elasticsearch/reference/master/search-aggregations-bucket-significanttext-aggregation.html

Fig. 5. TagSpheres: the context of the word secret without stopwords in the nine editions of William Preston's *Illustrations of Masonry*, 1775-1812.

5.3 Topic modeling

Topic modeling can help us find hidden semantic information about selected corpora. Statistical methods are used to discover the themes that are embedded in the texts and to reveal the connections of these themes and their changes over time (Blei et al., 2003). The AVOBMAT has an in-browser Latent Dirichlet Allocation function that draws on the jsLDA library[10] to calculate and graphically represent topic models. The LDA gives a predefined number of latent topics where each document may be viewed as a mixture of these topics. Besides the topic analysis AVOBMAT can also visualize the results of the modeling in various ways. As an improvement over the standard jsLDA tool, AVOBMAT allows for the adjustment of the LDA hyperparameters α and β that control the topic diversity of the documents and words. Bibliographic in-

[10] https://mimno.infosci.cornell.edu/jsLDA/

formation (i. e. authors, publication titles, dates) that plays a crucial role in knowledge discovery is hardly reflected in topic modelling. Unlike jsLDA, when listing topic documents, AVOBMAT displays the afore-mentioned basic bibliographic data, along with the probabilistic values for each document. It also presents the evolution of the different topics in an interactive time series. Moreover, before running the topic modelling, the user can filter and categorize the text corpus with the help of the metadata (currently 97 optional fields) associated with each document. The results can be exported in six different ways: document topics, topic words, topic summaries, topic-topic connections, doc-topic graph file (for Gephi) and complete sampling state (docID, word and topic number).

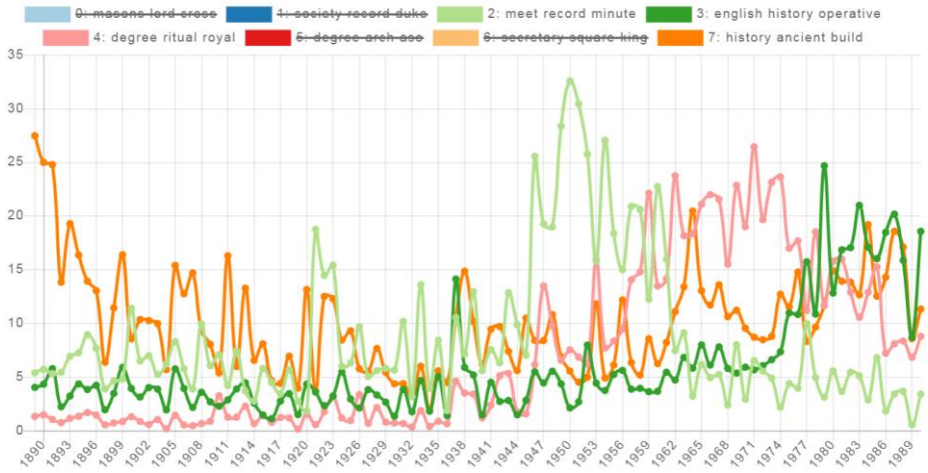


Fig. 6. Time series of topics with lemmatized words in the *Ars Quatuor Coronatorum* journal between 1890 and 1990. For instance, topic 4 (degree, ritual, royal, arch [Royal Arch: name of the 4th degree [rank] in masonic ritual hierarchy], ceremony. lecture, question] is clearly related to ritual studies in the journal. Topic 2 (meet, record, minute, pay, masons, elect, secretary apprentice) is concerned with the lodge meetings of Freemasons as recorded in their minutes. Topic 3 is about the history of operative (stonemason) Freemasonry.

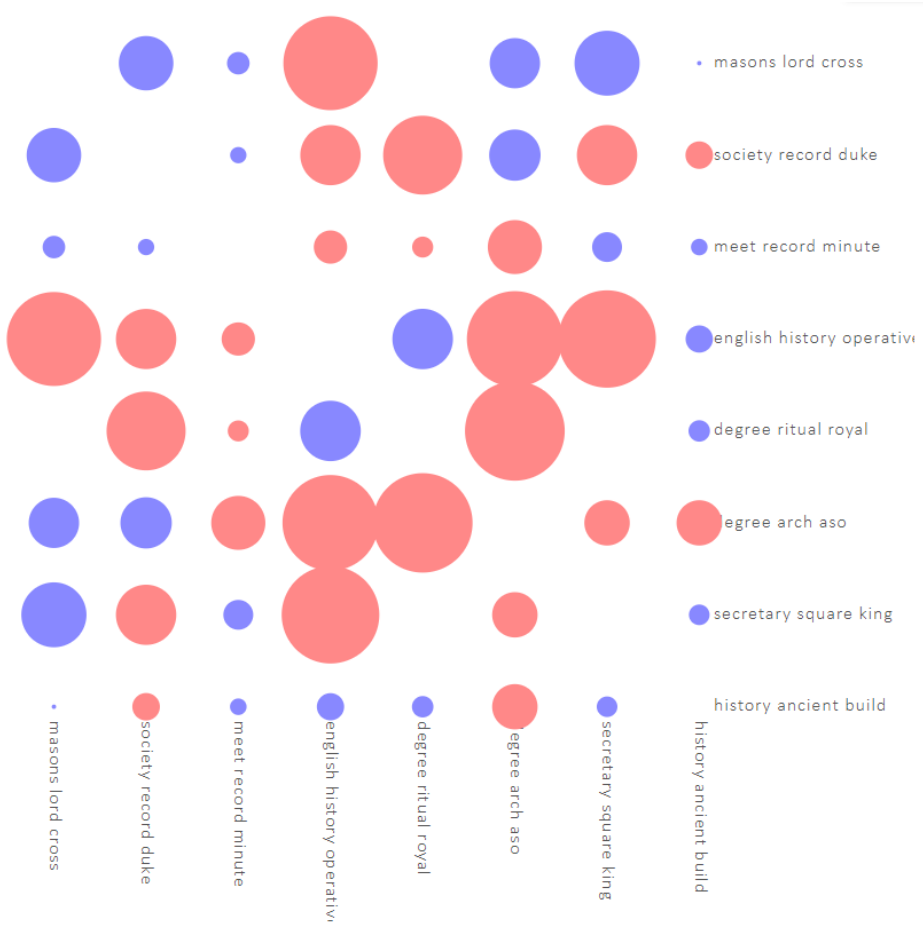


Fig. 7. Topic correlations in the *Ars Quatuor Coronatorum* journal.

5 Conclusion and future work

This paper has introduced the platform-independent and multilingual AVOBMAT system, which was primarily developed for scholars and students who are not familiar with command-line scripting. We have seen that this exploratory web application allows for a range of dynamic text and data mining tasks. The simple user-friendly web-based interface provides interactive parameter tuning and control from the pre-processing to the analytical stages.

The unique feature of the AVOBMAT toolkit is that it combines cutting-edge bibliographic data and computational text analysis research. It allows users to filter the uploaded datasets by metadata and full-text searches of various types and perform all the bibliographic, network and natural language analyses on the filtered datasets. In this way researchers can easily and interactively experiment with the different metada-

ta and content analyses, the parameters of which can be set online. Hence, it helps users realize the epistemological challenges, limitations and strengths of computational text analysis and visual representation of digital texts and datasets.

Besides revealing hitherto unknown connections of the different metadata fields and highlighting overlooked trends over time, the bibliographic data analysis can also highlight selection biases, errors in the bibliographic (meta)data (e.g. incorrect classifications) and reveal missing values and gaps in the data. Most data providers including libraries and profit-oriented companies have either not realized these, or if they did, they are reluctant to make this information public. Identifying these shortcomings helps researchers make informed decisions about their projects and critically analyse their datasets. It can also assist librarians in identifying the historical development regarding the creation of bibliographic records and improving the quality of the metadata of the repositories.

By combining distant and close reading approaches in our analytical framework, researchers can identify new perspectives for bibliographic data and textual analysis, discover novel insights, hidden patterns, themes and trends in digital collections. For instance, the use of bibliographic data allows scholars to perform diachronic topic analysis revealing wider semantic patterns in language use than a close reading of massive digital collections would provide. With the help of the bibliographic metadata we can filter our text corpora for advanced and comprehensive content analyses and carry out network analysis of the different metadata fields. The traditional contextual close reading examination is fostered by the TagSphere visualization tool and the keyword in context (KWIC) representation of the search queries.

The AVOBMAT application will be made available for the public in 2020 with some restrictions concerning the size of the digital collections to be uploaded and explored due to limited server capacities.

We plan several developments in the near future. For instance, we intend to extend the AVOBMAT multilingual system with named entity recognition (with disambiguation), parts of speech tagger, lexical richness and sentiment analysis tools. We would also like to increase the number of supported languages for lemmatization as well as the input file types. If server capacities permit, users will be allowed to upload their own pre-trained models.

Acknowledgements

The research was supported by the EU-funded Hungarian grant EFOP-3.6.1-16-2016-00008. It was also supported by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002. The project has been supported by the European Union and co-funded by the European Social Fund. We are grateful to Károly Kokas, Ákos Sándor, Gyula Nagy and Zoltán Erdődi of the Klebelsberg University Library for providing access to library repositories as well as their professional and technical support. Our thanks are due to Tamás Ficand, Gábor Simon and Gergely Dér who participated in the devel-

opment of the previous versions of the AVOBMAT. Simon and Dér devoted their BSc and MSc theses to this topic.

Bibliography

Blei, D. M., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning. Research. 3, 993–1022 (2003).

Fenlon, K. et al.: Tooling the Aggregator's Workbench: Metadata Visualization Through Statistical Text Analysis. Proceedings of the American Society for Information Science and Technology. 49, 1, 1–10 (2012).

Graham, S. et al.: Exploring Big Historical Data: the Historian's Macroscope. Imperial College Press, London (2016).

Hill, M. J. et al.: Reconstructing Intellectual Networks: From the ESTC's Bibliographic Metadata to Historical Material. In: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference: Copenhagen, Denmark, March 5–8, 2019. pp. 201–219. CEUR-WS.org (2019).

Jänicke, S., Scheuermann, G.: On the Visualization of Hierarchical Relations and Tree Structures with TagSpheres. In: Braz, J. et al. (eds.) Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 199–219. Springer International Publishing, Cham (2017).

Jockers, M. L.: Macroanalysis: Digital Methods and Literary History. University of Illinois Press., Champaign, IL. (2013).

Juršic, M., et al. Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules. Journal of Universal Computer Science 16, 9, 1190-1214 (2010).

Kleinman, S., LeBlanc, M.D., Drout, M., and Feng, W. (2019). Lexos. v4.0 https://github.com/WheatonCS/Lexos/.

Lahti, L. et al.: Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly. 57, 1, 5–23 (2019).

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Mahmoudi, M.R., Abbasalizadeh, A.: How Statistics and Text Mining can be Applied to Literary Studies? Digital Scholarship in the Humanities. 34, 3, 536–541 (2019).

Moretti, F.: Distant Reading. Verso, London; New York (2013).

Moreux, J.-P.: Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment. In: IFLA News Media Section, Lexington, United States (2016).

Péter, R.: Researching (British Digital) Press Archives with New Quantitative Methods. Hungarian Journal for English and American Studies. 17, 2, 283-300 (2011).

Péter, R.: Digitális és módszertani fordulat a sajtókutatásban: A 17–18. századi magyar vonatkozású angol újságcikkek 'távolságtartó olvasása' [Digital and Methodological Turn in the Study of the Press: a 'Distant Reading' of 17th-18th c. English Newspapers Concerning Hungary]. Aetas 29, 1, 5-30 (2015).

Prescott, A.: Bibliographic Records as Humanities Big Data. In: 2013 IEEE International Conference on Big Data. pp. 55–58 (2013).

Prescott, A.: Big Data in the Arts and Humanities: Some Arts and Humanities Research Council Projects. University of Glasgow, Glasgow (2015).

Schiuma, G., Carlucci, D.: Big Data in the Arts and Humanities: Theory and Practice. Taylor and Francis, Boca Raton, FL (2018).

Scrivner, O. et al.: Building Customized Text Mining Tools via Shiny Framework: The Future of Data Visualization. In: MAICS. (2017).

Scrivner, O., Davis, J.: Interactive Text Mining Suite: Data Visualization for Literary Studies. In: CDH@TLT. (2017).

Scrivner, O., Davis, J.: Topic Modeling of Scholarly Articles: Interactive Text Mining Suite. Presented at the Computational Linguistics and Intellectual Technologies Conference, Moscow, June 1-4, 2016 (2016).

Sinclair, S., and Rockwell, G. (2019). Voyant Tools. Web. http://voyant-tools.org/.

Tolonen, M. et al.: A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800. Liber quarterly. 25, 2, 87–116 (2015).

Varlamis, I., Tsatsaronis, G.: Visualizing Bibliographic Databases as Graphs and Mining Potential Research Synergies. In: 2011 International Conference on Advances in Social Networks Analysis and Mining. pp. 53–60 (2011).

Beszédtechnológia I.

Depresszió detektálása korrelációs struktúrán alkalmazott konvolúciós hálók segítségével

Jenei Attila Zoltán[1], Kiss Gábor[2]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
ja1504@hszk.bme.hu
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
kiss.gabor@tmit.bme.hu

Kivonat: Jelen kutatásban a depressziós állapot automatikus detektálásának lehetőségét vizsgáltuk a beszédjelből kinyert speciális korrelációs struktúrán alkalmazott konvolúciós neurális hálok segítségével. A depresszió korunk egyik legelterjedtebb gyógyítható pszichiátriai betegsége. A depressziótól szenvedő egyén életminőségét nagymértékben befolyásolja a depresszió súlyossága, ami extrém esetben öngyilkossághoz is vezethet. Ezek alapján kulcsfontosságú, hogy már korai stádiumában felismerhető legyen a betegség és az illető megfelelő kezelésben részesüljön, azonban a depresszió diagnosztizálása szakértelmet kíván, emiatt fontos a depresszió esetleges jelenlétének automatikus jelzése. Ebben a cikkben egy olyan eljárást mutatunk be, ami beszédjel feldolgozása alapján tisztán spektrális jellemzőkön keresztül képes felismerni a depressziót konvolúciós neurális hálók alkalmazásának segítségével. Bemutatjuk, hogyan változik a depresszió detektálásának pontossága különböző akusztikai-fonetikai jellemzők felhasználása alapján, illetve a korrelációs struktúrának változtatása következtében. A módszer alkalmazásával 84%-os pontossággal tudtuk elkülöníteni az egészséges és depressziós személyeket a beszédmintáik alapján.

1 Bevezetés

Számos betegség hatással bír a beszédkeltés folyamatára, és ez által hatással bír a kialakult beszédproduktumra is. Az egyes betegségek beszédre gyakorolt hatását lehetséges kimutatni a kialakult beszédproduktum akusztikai-fonetikai jellemzőinek megmérésével és ez alapján lehetséges az adott betegségek automatikus detektálása, illetve ezáltal szokás a beszédre mint egy objektív biomarkerre tekinteni (Sztahó és mtsai, 2019; Sztahó és mtsai, 2018; Liu és mtsai, 2017; Kiss és mtsai, 2017a; Tóth és mtsai, 2015; Orozco-Arroyave és mtsai, 2015; Cummins és mtsai, 2015).

A depresszió korunk egyik legelterjedtebb gyógyítható pszichiátriai betegsége (Friedrich, 2017), ami a WHO (World Health Organization) 2012-es felmérése alapján a harmadik leggyakoribb betegség világszerte (Marcus és mtsai, 2012).

A depresszió kialakulásának pontos okai még nem ismertek, azonban a depresszió élettani tünete leginkább a kortikális limbikus rendszer egyfajta diszfunkciójaként jelentkezik (Deckersbach és mtsai, 2006; Nestler és mtsai, 2002).

Depressziós állapot hatására az ettől szenvedő egyénnek a depresszió súlyosságának függvényében nehezére eshet elvégezni a napi teendőit, ami jelentős gazdasági károkat okozhat (Olesen és mtsai, 2012), ezen felül a depresszió súlyosbodásával megnövekedhet az öngyilkosság kockázata is (Hawton és mtsai, 2013). Azonban a depresszió diagnosztizálása szaktudást igényel, emiatt különösen fontos minden olyan megoldás ami segíthet a depresszió diagnosztizálásának támogatásában, illetve alkalmas lehet a depresszió veszélyének jelzésére.

A tény, hogy depresszió hatására megváltozik az emberi beszédproduktum már 1921-ben publikálta Kraeplin (Kraepelin, 1921), azonban a depressziós állapot és a beszéd kapcsolatának mélyebb vizsgálata, illetve a depresszió automatikus detektálása a megváltozott beszédproduktum alapján újszerű kutatási területnek számít, amit elsősorban az egyre nagyobb depressziós beszédadatbázisok megjelenése, illetve az informatika fejlődése tett lehetővé (Cummins és mtsai, 2015).

A korábbi kutatások esetében a depressziós és egészséges személyek automatikus elkülönítését, függetlenül a depressziós állapot súlyosságától, 50-86% közötti pontossággal voltak képesek megvalósítani beszédjel feldolgozás alapján (Kiss és Vicsi, 2017b; Kiss és Vicsi, 2017c; Alghowinem és mtsai, 2013; Ooi és mtsai, 2013; Cummins és mstai, 2013; Low és mstai, 2009). Természetesen a depresszió felismerésének a pontossága az egyes kutatások esetében nagyban függhet a kutatásban használt adatbázistól, illetve az adatbázis feldolgozottságától, az alkalmazott módszerektől. Annak ellenére, hogy több kutatás bizonyította már, hogy lehetséges a depresszió automatikus detektálása beszédjel feldolgozás alapján, több nyitott kérdés is van még, úgy, mint mely akusztikai-fonetikai jellemzők a legalkalmasabbak a depressziós állapot detektálásához, illetve milyen szintű feldolgozás szükséges a depressziós állapot detektálásához a minél nagyobb pontosság elérése érdekében.

Az utóbbi időben számos tanulmány a beszédakusztikai jellemzők széles skáláját alkalmazta annak érdekében, hogy (főleg bináris) osztályozást végezzen a depressziós és egészséges alanyok elkülönítésére (Kiss és Vicsi, 2017b; Vlasenko és mtsai, 2017; Cummins és mtsai, 2015; Valstar és mtsai, 2013). Azonban a megfelelő akusztikai-fonetikai jellemző halmaz kiválasztást nehezíti, hogy az eddig rendelkezésre álló depressziós beszédadatbázisok csupán 50-150 főtől tartalmaznak beszédmintákat (Cummins és mtsai, 2015; Kiss és Vicsi, 2017b), így az alkalmazott gépitanuló eljárások értelemszerűen csak limitált méretű jellemzővektorral képesek dolgozni a túltanulás elkerülése végett.

Jelen kutatásban Williamson és mtsai. 2013-ban publikált korrelációs mátrix alapú megoldását használjuk fel (Williamson és mtsai, 2013). Az adott publikációban az alacsony szintű akusztikai-fonetikai jellemzők auto- és keresztkorrelációs struktúrája alapján, nagy pontossággal képesek voltak a depressziós állapot súlyosságát becsülni. A magas pontosság mellett még figyelemre méltó volt a kutatásban, hogy mindösszesen a beszédjelből kinyert MFCC (Mel-frequency cepstral coefficients) és a formáns frekvenciák jellemzőkre támaszkodtak. Az eredményeket a német nyelvű depressziós beszédadatbázison érték el (Valstar és mtsai, 2013). Az eljárást már korábban sikeresen alkalmazták az agyi EEG (Electroencephalography) jeleken a kezdődő epilepszia jelzésére (Williamson és mtsai, 2011). Az auto- és keresztkorrelációs eljárás egy adott jellemzővektor halmazból kiindulva előállítja annak egy speciális korrelációs mátrix struktúrájú reprezentációját. A mátrix egyes celláiban a jellemzővektor halmazból vett két jellemzővektor (a két jellemzővektor lehet ugyanaz) korrelációs együttható értéke

található, meghatározott eltolások mellett. Az eljárás pontos ismertetését a 3.2-es fejezetben részletezzük.

Az előállított korrelációs mátrix az átlóra szimmetrikus, emiatt Williamson és mtsai. az előállított korrelációs mátrix sajátértékeit számíttatták ki, majd azoknak csupán egy részhalmazát használták fel a gépi tanuló eljárás bemeneteként, és becsülték ez alapján depresszió súlyosságát.

Az eljárás sikeressége feltehetőleg abban rejlik, hogy a beszéd nagy időablakban vett megváltozott struktúráját képes megfelelően reprezentálni. Korábbi kutatásunkban a korrelációs mátrix alapú eljárást mi is sikeresen alkalmaztuk egyidejűleg több betegség felismerésére (Sztahó és mtsai, 2018), ahol is 3 különböző betegség (Parkinson kór, depresszió és egyéb gégészeti elváltozások) és egészséges beszélőktől származó beszédminták automatikus elkülönítését végeztük el 78%-os pontossággal. Azonban Williamson és mtsai. által publikált eljárásnak van egy fő hátránya, ugyanis az előállított korrelációs mátrix sajátértékekkel vett reprezentációja nem feltétlenül optimális és még mindig redundáns, illetve túl nagy méretű. Emiatt szükséges a sajátértékek alapú reprezentáció dimenziójának további csökkentése, aminek megfelelő megválasztásától a gépi tanuló eljárás pontossága és általánosító képessége is nagyban függhet. A problémát tovább rontja, ha egyszerre sok akusztikai-fonetikai jellemzőt is fel szeretnénk használni a depresszió felismerésére.

Emiatt jelen kutatásban a korrelációs mátrixok sajátértékeinek használata helyett, a mátrixokat közvetlenül alkalmaztuk a gépi tanuló eljárás bemeneteként. Ehhez konvolúciós (CNN) neurális hálókat alkalmaztunk. Az eljárás előnye, hogy így a gépi tanuló eljárás feladata a korrelációs mátrix megfelelő feldolgozása is. A kutatásban még újszerű, hogy nemzetközi viszonylatban is nagy mintaszámúnak számító, közel 200 beszédmintán tudtuk tesztelni az eljárást.

A cikk a következő felépítést követi. A bevezetés után a második fejezetben bemutatjuk a felhasznált beszédadatbázist. A harmadik fejezetben bemutatjuk az alkalmazott alacsony szintű jellemzőket, a korrelációs mátrix kiszámításának módját és az azon alkalmazott konvolúciós neurális hálók felépítését, illetve a kiértékelési módszereket. A negyedik fejezetben bemutatjuk az eredményeket. Az ötödik fejezetben röviden összefoglaljuk a kutatás fő eredményeit és a további terveinket.

2 Magyar Depressziós Beszédadatbázis

A kutatás során a Magyar Depressziós Beszédadatbázis beszédmintáira támaszkodtunk. A beszédadatbázist folyamatosan bővítjük. Jelen kutatásban a beszédadatbázisban elérhető 91 egészséges és 91 depressziós személytől gyűjtött beszédmintákat használtuk fel, minden személytől pontosan egy beszédmintát. Az egészséges személyek esetén csak olyan személyek beszédmintáit használtuk fel, akik - saját bevallásuk alapján - nem szenvedtek semmilyen olyan betegségtől, ami hatással bírhat a beszédükre. A depresszióval diagnosztizált betegek esetén szintén csak az olyan személyektől származó beszédmintákat használtunk fel, akik nem voltak diagnosztizálva más olyan betegséggel, ami szintén hatással bírhat a beszédproduktumukra (pl.: Parkinson kór, ALS).

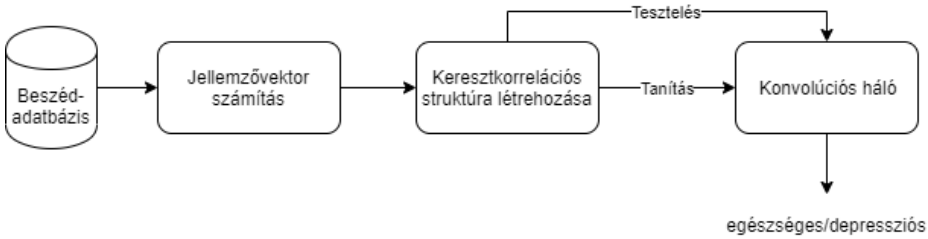
A beszédminták gyűjtését a Semmelweis Egyetem Pszichiátriai és Pszichoterápiás Klinikájával együtt végezzük. A beszédminták gyűjtésénél törekedtünk arra, hogy a

beszélők lefedjék a depresszió súlyosságának különböző fokozatait, az egészséges állapottól az egészen súlyos depresszióig. A depressziós személyek esetében körülbelül egyenletes eloszlással szerepeltek alanyok a BDI-II (Beck Depression Inventory-II) (Beck és mtsai, 1996) által definiált depressziós súlyosság szerinti kategóriák között, úgy, mint az enyhe depresszió, közepes depresszió és súlyos depresszió.

A vizsgált személyeknek egy fonetikusan gazdag mesét ("Az északi szél és a Nap") kellett felolvasniuk. A felvételek csendes helyiségben kerültek rögzítésre 44,1 kHz mintavételi frekvenciával, csiptetős mikrofonnal.

3 Módszerek

A kutatásunkban alkalmazott, a depressziós állapot detektálására alkalmas módszer folyamatát az 1. ábra mutatja be. Az eljárás bemenetén a beszédminta áll, míg az eljárás kimenete a bemondó bináris osztályozása (egészséges/depressziós) a beszédmintája alapján. Az eljárás először az adott beszédmintából különböző akusztikai-fonetikai jellemzőenergia vektorokat nyer ki. Ezt követően a kiszámított jellemzővektorok részhalmazából előállítja azok auto- és keresztkorrelációs mátrixát. A létrehozott kétdimenziós korrelációs mátrix lesz a bemenete a 2D konvolúciós hálónak, ami tanítás esetén létrehozza a megfelelő modellt, majd tesztelés esetén a modell segítségével elvégzi a bemondó bináris osztályozását, ami az eljárás végső kimenete.



1. ábra. A kutatásban bemutatott depressziós állapot detektálására alkalmas módszer folyamat ábrája.

3.1 Felhasznált akusztikai-fonetikai jellemzők

Az akusztikai-fonetikai jellemzők számítása előtt a beszédminta minden esetben csúcsértékre lett normalizálva, ezzel kiküszöbölve a felvételek rögzítése esetén esetlegesen felmerülő eltérő erősítésbeli különbségeket. A következő alacsony szintű jellemzőket használtuk:

Mel-sávos energiaértékek: Az emberi hallás frekvenciabeli felbontásához hasonló sávokban adja meg az energiaértékeket. A sávokat 100-dik mel-től kezdve 100 melenkénti összegzéssel valósítottuk meg, összesen 27 mel-sávos energiaérték kiszámíttatásával, amivel körülbelül 60 Hz és 8 kHz között végeztük el a beszédjel frekvenciabeli felbontását.

MFCC együtthatók: Az MFCC együtthatók alkalmazása és azoknak fontossága a beszédjel feldolgozás területén bevett gyakorlatnak számít. Az MFCC együtthatókat a 27 mel-sávos energiaérték diszkrét koszinusz transzformáltjaként számíttattuk ki és összesen 14 együtthatót használtunk fel végül.

Formáns frekvenciák: Formáns frekvenciákon a beszédjel feldolgozás esetében, a rezonátorüregek által felerősített felhangnyalábok burkoló görbéinek maximum helyeit értjük. A kutatás során az első három formánsfrekvencia értékeket számíttattuk ki és használtunk fel, amikre a továbbiakban mint F1, F2 és F3 hivatkozunk.

Formáns frekvenciák sávszélessége: Az adott formánsfrekvencia sávszélessége alatt, a formánsfrekvencia 3 dB-es csökkenésénél mért sávszélességet értjük. A kutatás során az F1, F2 és F3 formáns frekvenciák sávszélességét számíttattuk ki és használtuk fel, amiket a továbbiakban B1, B2 és B3-al jelölünk.

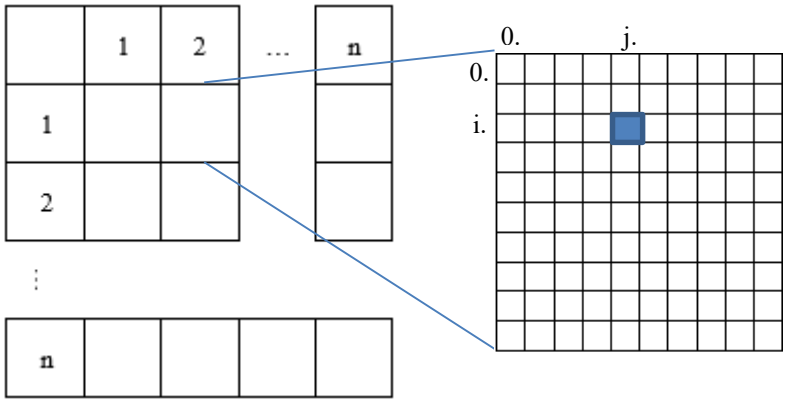
A kutatás során felhasznált akusztikai-fonetikai jellemzőket 10 ms-os lépésközzel, 50 ms-os ablakkal számíttattuk ki. A mel-sávos energiaértékeket és az MFCC együtthatókat a teljes beszédmintából, míg a formáns frekvenciákat és azok sávszélességeit pedig a zöngés szakaszokból számíttattuk ki. Így minden egyes akusztikai-fonetikai jellemző esetében egy jellemzővektort kaptunk az adott beszédmintára, ahol is a mel-sávos energiaértékeket és az MFCC együtthatókat tartalmazó vektorok hossza, illetve a formáns frekvenciák és azok sávszélességeinek hossza megegyezik.

A jellemzők kiszámításhoz a Praat szoftvert használtuk (Boersma, 2001).

3.2 Korrelációs struktúra

A jellemzővektorok adott halmazaiból azok auto- és keresztkorrelációs együtthatóit tartalmazó korrelációs mátrixait hoztuk létre. A korrelációs mátrix számítását tömören ismertetjük, bővebb leírása Williamson és mtsai 2013-as cikkében található (Williamson és mtsai, 2013)

A korrelációs mátrix felépítését a **2. ábra** szemlélteti, ahol *n* jelöli a korrelációs mátrix bemeneti jellemzővektorainak számát.



2. ábra. A korrelációs mátrix felépítése.

A korrelációs mátrix $n \cdot n$ darab almátrixból épül fel (2. ábra bal oldala). A főátló mentén található almátrixok az adott jellemzővektorok autokorrelációs együtthatóit, míg a többi almátrix két különböző jellemzővektor keresztkorrelációs együtthatóit tartalmazza *dt* darab eltolás mellett. Jelen kutatás során a $dt = 10$ értéket alkalmaztunk.

Minden almátrix összesen $dt \cdot dt$ darab korrelációs együtthatót tartalmaz. (Vagyis a teljes mátrxinak összesen $(n \cdot dt) \cdot (n \cdot dt)$ darab cellája van.) Az adott almátrix egyértelműen meghatározza, hogy mely két jellemzővektor korrelációs értékei találhatók benne a felhasznált jellemzővektor halmazból. Az almátrix első cellája (0. sor, 0. oszlop) a két jellemzővektor eltolás nélküli korrelációs együttható értékét tartalmazza. Az adott almátrix egy tetszőleges i. sorában és j. oszlopában található korrelációs együttható értéke pedig az első jellemzővektor *i* szer vett eltolása és a második jellemzővektor *j* szer vett eltolása melletti korrelációs együttható értékét tartalmazza (2. ábra jobb oldala). A korrelációs mátrix felépítéséből fakadóan a fő átló elemei csupa 1-et tartalmaznak, illetve a mátrix szimmetrikus a fő átlóra. Fontos megjegyezni, hogy értelemszerűen komolyabb módosítások nélkül, csak egyforma hosszúságú jellemzővektorokra működik az eljárás. Kutatás során az eltolás mértékére 3 különböző értéket is kipróbáltunk (1, 2 és 8), vagyis például ha 2 volt az eltolás mértéke és az *i* értéke éppen 3 volt, akkor az adott jellemzővektort 6 értékkel toltuk el.

Az eljárás alapján minden egyes beszédmintából pontosan egy korrelációs mátrixot számíttattunk ki egy adott jellemzővektor halmaz esetében.

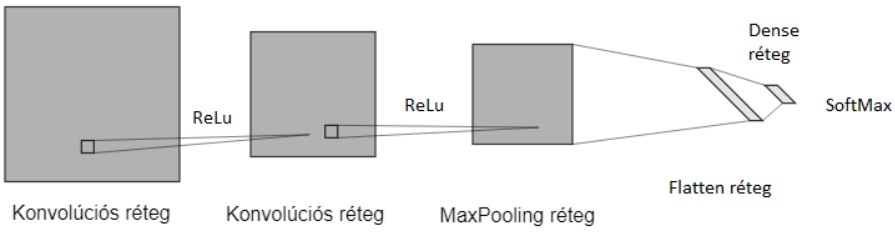
Összesen 5 különböző jellemzővektor halmazt alkalmaztunk a kutatás során:

- Mel-sávos energiaértékeket tartalmazó jellemzővektorok
- MFCC együtthatókat tartalmazó jellemzővektorok
- Formáns frekvenciákat tartalmazó jellemzővektorok
- Formáns frekvenciák sávszélességét tartalmazó jellemzővektorok
- Formáns frekvenciák és azok sávszélességeit tartalmazó jellemzővektorok

3.3 Konvolúciós háló

Az osztályozó algoritmusnak 2D konvolúciós neurális hálókat alkalmaztunk. A gépi tanuló eljárás bemenete az adott jellemzővektor halmazból kiszámított korrelációs mátrix volt.

Az algoritmus létrehozása Phyton kódban történt TensorFlow környezetben. Felépítését a **3. ábra** szemlélteti.



3. ábra. Az alkalmazott konvolúciós háló szerkezete.

A konvolúciós rétegek 32 filtert használtak, amik mérete $10 \cdot 10$-es az almátrix méretének megfelelően. A kernel lépésközének szintén 10 volt beállítva az eltolások száma (*dt*) alapján. A maxPooling kernel mérete $2 \cdot 2$-es volt és same paddinget alkalmaztunk. Az első három réteg után dropOut regulációt használtunk, ami véletlenszerűen a neuronok 25 %-át figyelmen kívül hagyta a tanítás során. A Flatten rétegbe már a 32 filter értéke került, így mérete $1 \cdot 32$-es volt, ami a bemenete egy fully connected neurális hálónak (a Dense rétegnek). Ennek kimenetén SoftMax függvényt alkalmaztunk bináris osztályozáshoz.

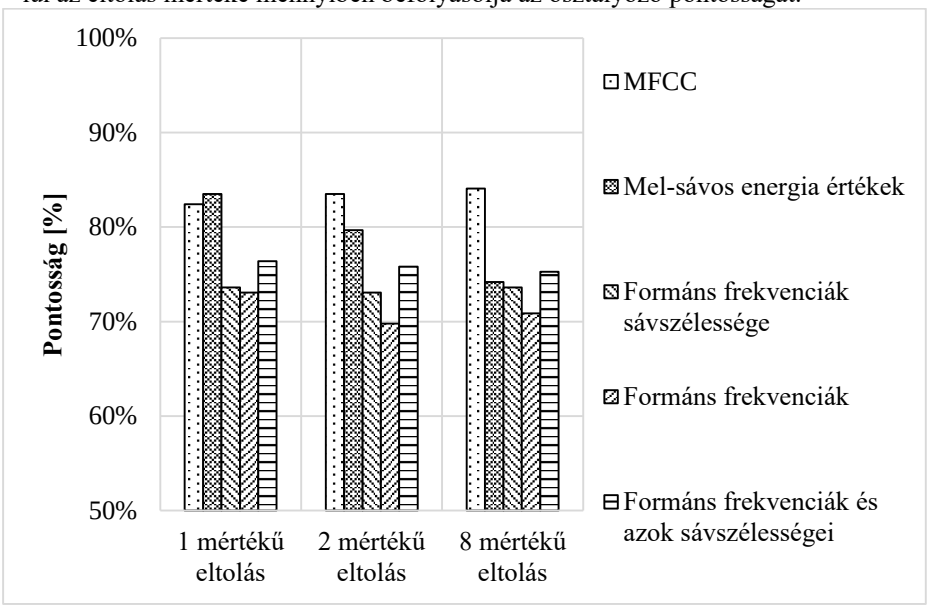
3.4. Kiértékelési módszerek

A viszonylag alacsony mintaszám miatt az ilyenkor szokásos teljes kereszt-validációs eljárást alkalmaztunk, a háló tanítás és tesztelése során. Vagyis minden egyes mintát egyszer, mint teszt halmaz a maradék mintákat pedig mint tanító halmaz alkalmaztuk. Fontos megjegyezni, hogy természetesen ez által a tesztelő és tanító halmazok minden esetben diszjunktak voltak.

Az osztályozási kísérletek során a minél nagyobb pontosság (helyesen osztályozott minták száma osztva az összes minta számával) elérését tűztük ki célul. Emellett vizsgáltuk még az orvosi diagnosztikában nagy fontosságú és emiatt gyakran alkalmazott specificitás és szenzitivitás értékeket is, hiszen a gyakorlatban nem feltétlenül számít ugyanakkora hibának, ha egy egészséges embert depressziósnak ítélünk, mint fordítva.

4. Eredmények

Összesen 15 különböző osztályozási kísérletet valósítottunk meg. 5 jellemzővektor halmazt vizsgáltunk és mindegyikből 3 különböző mértékű eltolás alkalmazásával állítottunk elő korrelációs mátrixokat (lásd 3.2-es fejezet). Elsősorban azt vizsgáltuk, hogy mely jellemzővektor halmazzal lehet elérni a legnagyobb pontosságot, illetve azon belül az eltolás mértéke mennyiben befolyásolja az osztályozó pontosságát.



4. ábra. Depresszió felismerésének pontossága különböző jellemzővektor halmazokból és eltolás mértékkel előállított korrelációs mátrixok alapján.

4. ábrán látható a 15 különböző kísérlet elvégzése során kapott pontosságok értéke. Amint látható MFCC együtthatók felhasználásával 8 mértékű eltolással előállított korrelációs mátrix esetén kaptuk a legnagyobb pontosságot, ebben az esetben 84%-os pontossággal tudtuk elkülöníteni a depressziós és egészséges beszélőket. Továbbá megfigyelhető, hogy az eltolás mértékének a növelésével a legtöbb esetben csökkenő pontosság értékeket kaptunk, kivétel az MFCC és a formáns frekvenciák esetében. Ezért további vizsgálatokat végeztünk 16 és 32 mértékű eltolást alkalmazva, ahol jelentősebb csökkenést tapasztaltunk a pontosságban.

Az 1. táblázatban látható minden egyes kísérlet esetében az elért pontosság, specificitás, és szenzitivitás értékek. Félkövérrel kiemeltük az egyes metrikák szerinti legnagyobb elért értékeket. Megfigyelhető, hogy minden esetben ezeket a maximális értékeket az MFCC jellemzővektor halmaz használata esetében kaptuk (1. táblázat). Továbbá megfigyelhető, hogy a formáns frekvenciák és azok sávszélességei együttes felhasználása javította az osztályozás pontosságát 73,6%-ról 76,4%-ra, azonban az így elért pontosság elmarad az MFCC-vel (82,4% - 84,1%) és a mel-sávos energiaértékekkel (74,2% - 83,5%) elért pontosságoktól.

A legjobb eredményünket (84,1%) összehasonlítva hasonló kutatások eredményeivel (50-86%) kijelenthető, hogy viszonylag nagy pontosságot voltunk képesek elérni, de természetesen, ahogy arra már a bevezetőben utaltunk, az eredmények nehezen összehasonlíthatók. Legpontosabb összehasonlítást a Magyar Depressziós Beszédadatbázison általunk publikált korábbi eredményekkel lehetséges megtenni, ahol is eltérő módszereket alkalmazva 83%-os (Kiss és Vicsi, 2017c) illetve 86%-os (Kiss és Vicsi, 2017b) pontosságot tudtunk elérni.

1. táblázat: Depresszió felismerésének leíró jellemzői az eltérő korrelációs mátrixok alapján

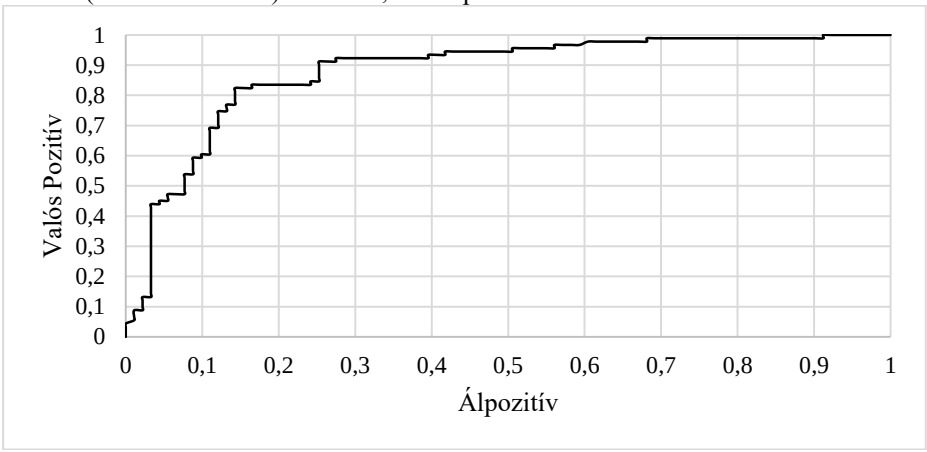
Jellemzővektor halmaz neve	Eltolás mértéke	Specificitás	Szenzitivitás	Pontosság
MFCC	1	81,3%	**83,5%**	82,4%
	2	83,5%	**83,5%**	83,5%
	8	**87,9%**	80,2%	**84,1%**
Mel-sávos energia értékek	1	84,6%	82,4%	83,5%
	2	80,2%	79,1%	79,7%
	8	73,6%	74,7%	74,2%
Formáns frekvenciák	1	76,9%	69,2%	73,1%
	2	73,6%	65,9%	69,8%
	8	70,3%	71,4%	70,9%
Formáns frekvenciák sávszélessége	1	78,0%	69,2%	73,6%
	2	82,4%	63,7%	73,1%
	8	78,0%	69,2%	73,6%
Formáns frekvenciák és azok sávszélességei	1	76,9%	75,8%	76,4%
	2	75,8%	75,8%	75,8%
	8	75,8%	74,7%	75,3%

Ez alapján megállapítható, hogy körülbelül ugyanakkora pontossággal voltunk képesek detektálni a depressziót mint korábban. Azonban fontos megjegyezni, hogy az általunk most bemutatott eljárás független a beszélő nemétől, és nem igényli a beszédminták bonyolult, beszédhang szintű előfeldolgozását, szemben a korábbi munkáinkkal. Viszont a korábbi eredményeink a Magyar Depressziós Beszédadatbázis egy régebbi állapotán készültek, ahol összesen csak körülbelül 130 beszélőtől állt rendelkezésünkre beszédminta. Összeségében kijelenthető, hogy az eredmények bizakodásra adnak okot és a bemutatott módszer további vizsgálata mindenképpen fontos.

Egy valós rendszer esetében a fő cél, hogy ha valaki depressziós azt a rendszer helyesen depressziósnak jelezze (valós pozitív arány), míg az kisebb hiba, hogyha valakit egészségesként dönt depressziósnak (álnegatív). Emiatt a legjobb pontosságot elérő beállítások mellett megvizsgáltuk, hogy egy adott valós pozitív arány mellett mekkora

lenne az álnegatívok aránya, amit a ROC (reciver operating characteristic) görbe megadásával szemléltetünk (5. ábra).

Ennek megvalósítására a neurális háló közvetlen kimenete adott lehetőséget, hiszen valójában 0 és 1 közötti számot adott vissza, ahol 0,5 értéknél kisebbek jelentették az egészséges osztályozást, míg az ennél nagyobbak a depressziós osztályozást. Így a komparátor értékét változtatva 0 és 1 között megfigyelhető, hogy adott valós pozitív arány mellett, mekkora lenne az álnegatív arány. A ROC-görbe integrálása alapján az AUC (area under curve) értékre 0,79-t kaptunk.



5. ábra. A depresszió detektálásának ROC görbéje MFCC jellemzővektor halmazból 8 mértékű eltolással számított korrelációs mátrixok alapján.

A ROC-görbe alapján megállapítható (5. ábra), hogy például 90%-os valós pozitív arányt elvárva az álnegatívok aránya már 25%. A gyakorlatban egy önálló diagnosztikát támogató rendszernek valószínűleg ennél nagyobb pozitív arány mellett kisebb álnegatív arányt kellene biztosítani, ahhoz hogy igazán jól alkalmazható lehessen, emiatt kívánatos lenne a módszer további fejlesztése.

5. Összefoglalás

Jelen kutatásban a depressziós állapot automatikus detektálásának lehetőségét mutattuk be beszédjel feldolgozás alapján. A kutatás eredménye hozzájárulhat egy a depresszió diagnosztizálását támogató minél pontosabb rendszer megvalósításához. Egy ilyen esetleges rendszer megvalósítása nagyban segíthetné a depresszió meglétének automatikus felismerését. A figyelmeztetés alapján az esetlegesen depressziótól szenvedő alany minél hamarabb megfelelő szakemberhez fordulhatna segítségért, ami megnövelheti a gyógyulás esélyét, illetve csökkentheti a kezelés időtartamát is. Ezek pedig csökkentenék a depresszió által okozott gazdasági károkat, illetve az öngyilkosságok számát.

A kutatásban a depresszió detektálásának lehetőségét a mel-sávos energiaértékek, az MFCC együtthatók, a formáns frekvenciák és azok sávszélességei, mint alacsony szintű akusztikai-fonetikai jellemzőkre támaszkodva ismertettük. A bemutatott alacsony

szintű akusztikai-fonetikai jellemzők adott részhalmazaiból képeztük azoknak egy speciális korrelációs struktúráját (mátrixát), amit mint bemenet kapott meg egy konvolúciós neurális hálót megvalósító gépi tanuló eljárás. Megvizsgáltuk, hogy mely akusztikai-fonetikai jellemzőhalmazra támaszkodva, milyen korrelációs struktúra esetén mekkora pontossággal detektálható a depressziós állapot. A vizsgálatok alapján legjobb eredményt az MFCC együtthatókra támaszkodva értünk el, 8 mértékű eltolást alkalmazva a korrelációs mátrix kialakítása során (84%-os pontosság). Az eredményt összehasonlítva más kutatások hasonló eredményeivel (50% - 86%-os pontosság), kijelenthető, hogy magas pontossággal voltunk képesek a depresszió automatikus felismerésére beszédjel feldolgozás alapján.

Az általunk bemutatott módszernek számos előnye van. A fő előnyei közt említhető, hogy független a vizsgált személy nemétől, nem szükséges hozzá bonyolult előfeldolgozása a beszédmintának (például beszédhangszintű szegmentálása) és az általunk bemutatott eredményt képesek voltunk csupán a beszéd MFCC együtthatóira támaszkodva elérni. Fontos továbbá azt is megjegyezni, hogy az általunk alkalmazott módszerekkel minimális volt a túltanulás veszélye.

Jelen kutatást mindenképpen folytatni tervezzük. A jövőben több vizsgálatot is tervezünk megvalósítani. Az eljárást tesztelni fogjuk a tovább bővített Magyar Depreszszziós Beszédadatbázison (200-200 egészséges és depressziós beszélőtől gyűjtött mintaszám a cél). A konvolúciós háló struktúrájának módosításával lehetővé tenni, hogy az eljárás egyszerre több és újabb akusztikai-fonetikai jellemző halmazokból elállított korrelációs mátrixokat is képes legyen a bemenetén fogadni. Egyéb olyan jellemzők felhasználásának megvizsgálása (pl: prozódiai jellemzők), amiket bár nem lehetséges vagy érdemes felhasználni a korrelációs mátrix(ok) előállítása során, azonban értékük bizonyítottan megváltozik a depressziós állapot hatására, így hasznosak lehetnek a depressziós állapot detektálásában. Nemek szerint eltérő modellek alkalmazása esetében megvizsgálnánk, hogy az vajon javít-e az általunk bemutatott módszer pontosságán.

Köszönetnyilvánítás

Project no. K128568 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme.

Hivatkozások

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., Parker, G.: A comparative study of different classifiers for detecting depression from spontaneous speech. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8022-8026) (2013)

Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F.: Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. J. Pers. Assess. 67, 588–597. (1996)

Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345. (2001).

Cummins, N., Epps, J., Ambikairajah, E.: Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7542-7546). (2013)

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T. F.: A review of depression and suicide risk assessment using speech analysis. Speech Communication 71, pp. 10–49, (2015)

Deckersbach, T., Dougherty, D. D., Rauch, S. L.: Functional imaging of mood and anxiety disorders. Journal of Neuroimaging, 16(1), 1-10. (2006)

Friedrich, M. J.: Depression is the leading cause of disability around the world. Jama, 317(15), 1517-1517. (2017)

Hawton, K., i Comabella, C. C., Haw, C., Saunders, K.: Risk factors for suicide in individuals with depression: a systematic review. Journal of Affective Disorders, 147(1 3), 17-28. (2013)

Kiss, G., Simin, L., Vicsi, K.: Estimation of the severity of depression based on speech processing on Hungarian langugae (original title: Depresszió súlyosságának becslése beszédjel alapján magyar nyelven). In XIII. Magyar Számítógépes Nyelvészeti Konferencia,(MSZNY2017). Conference (pp. 125-135). (2017a)

Kiss, G., Vicsi, K.: Mono-and multi-lingual depression prediction based on speech processing. International Journal of Speech Technology, 20(4), 919-935. (2017b)

Kiss, G., Vicsi, K.: Comparison of read and spontaneous speech in case of automatic detection of depression. In 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (pp. 000213-000218). IEEE. (2017c)

Kraepelin, E.: Manic depressive insanity and paranoia. J. Nerv. Ment. Dis. 53, 350. (1921)

Liu, Y. , Lee, T.., Ching, P. C., Law, T. K. T., Lee., K. Y. S.: Acoustic assessment of disordered voice with continuous speech based on utterance-level ASR posterior features" in INTERSPEECH 2017 pp. 2680–2684. (2017)

Low, L. S. A., Maddage, N. C., Lech, M., Allen, N.: Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents. In 8th IEEE International Conference on Cognitive Informatics (pp. 346-350). (2009)

Marcus, M., Yasamy, M. T., van Ommeren, M., Chisholm, D., Saxena, S.: Depression: A global public health concern (www.who.int) (2012)

Nestler, E. J., Barrot, M., DiLeone, R. J., Eisch, A. J., Gold, S. J., Monteggia, L. M.: Neurobiology of depression. Neuron, 34(1), 13-25.7. (2002)

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., Jönsson, B., CDBE2010 Study Group, European Brain Council: The economic cost of brain disorders in Europe. European Journal of Neurology, 19(1), 155-162. (2012)

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J., Skodda, S., Rusz J., Nöth, E.: Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease. in INTERSPEECH 2015 pp. 95-99, Dresden, Germany, (2015)

Ooi, K. E. B., Lech, M., Allen, N. B.: Multichannel weighted speech classification system for prediction of major depression in adolescents. IEEE Transactions on Biomedical Engineering, 60(2), 497-506. (2013)

Sztahó, D., Kiss, G., Tulics, M. G., Vicsi, K.: Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features. In 2018 41st International Conference on Telecommunications and Signal Processing (TSP) (pp. 1-4). IEEE. (2018)

Sztahó, D.; Kiss, G.; Tulics, M. G.; Dér-Hajduska, B.; Vicsi, K.: Automatic discrimination of several types of speech pathologies. In: 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD 2019) Paper: 119 , 2 p.(2019)

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., Kálmán, J.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In Proceedings of INTERSPEECH 2015 (pp. 2694-2698) (2015)

Valstar, M. F., Schuller, B. W., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013: the continuous audio/visual emotion and depression recognition

challenge. in: 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM. pp. 3–10, (2013)

Vlasenko, B., Sagha, H., Cummins, N., Schuller, B.: Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. INTERSPEECH 2017, pp. 3266–3270, Stockholm, Sweden, (2017)

Williamson, J. R., Bliss, D. W., Browne, D. W.: Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial EEG. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 665-668). IEEE. (2011)

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., Mehta, D. D. Vocal biomarkers of depression based on motor incoordination. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 41-48. (2013)

Nagyszótáras beszédfelismerés morfémaalapú rekurrens nyelvi modell használatával

Grósz Tamás

Aalto University, Finland
tamas.grosz@aalto.fi

Kivonat A klasszikus beszédfelismerő rendszerek számára hatalmas kihívást jelentenek az agglutináló nyelvek, hiszen pontos eredmények eléréséhez hatalmas szótárakra van szükség a ragozás és a szóösszetétel miatt. A probléma főleg a nyelvi modell részét érinti a felismerőnek, tekintve, hogy túl nagy szótárméret esetén a tanulási fázis rendkívül nehéz, ez pedig szuboptimális modellhez vezethet. Ezen problémára megoldást jelenthet, ha szavak helyett azoknál kisebb egységet, morfémákat használunk a nyelvi modellezés során. A cikkben bemutatásra kerül egy morfémaalapú, rekurrens neuronhálós nyelvi modellt alkalmazó beszédfelismerő, amely használatával szignifikánsan jobb eredményeket tudtunk elérni egy magyar nyelvű beszédkorpuszon mint a hagyományos szószintű megközelítéssel.
Kulcsszavak: beszédfelismerés, nyelvi modell, morféma, rekurrens neuronháló

1. Bevezetés

Az elmúlt pár évben elfogadott ténnyé vált, hogy mély neuronhálós akusztikus és nyelvi modellekkel lehet elérni a legjobb beszédfelismerési pontosságot (Hinton és mtsai, 2012). Ezen új beszédfelismerő rendszerek többsége a nyelvi modell építése során szavakat használ építőelemként, ami angol nyelv esetén jól működik, azonban komoly problémát okoz agglutináló nyelvek esetében.

A legnagyobb problémát a szóalaki változatosság okozza, amely egy fontos jellemzője a morfológiailag gazdag nyelveknek. Sok szóalak esetén rendkívül nagy méretű szótárat kell használnunk, hogy elfogadható pontosságot tudjunk elérni, ez pedig megnehezíti a nyelvi modell tanítását, mivel nagy szótár esetén viszonylag kevés tanítóminta áll rendelkezésünkre osztályonként.

Megoldásként módosíthatjuk a nyelvi modellünket, hogy szavak helyett azoknál kisebb egységeket használjon. Egy ilyen lehetséges egység a morféma, amit korábban már sikeresen használtak finn és magyar nyelvű beszédfelismerőkben. Extrém esetben átválthatunk akár karakter szintű nyelvi modellre is, az ún. end-to-end beszédfelismerő rendszerek jelentős része ezt a megoldást használja. Mindkét megközelítés esetén számottevően csökken a szótárméret, ezáltal könnyebbé válik a nyelvi modell tanítása. Munkánkban mi a morfémaalapú megközelítést vizsgáltuk.

Cikkünkben egy általános módszert mutatunk be, amelynek segítségével morfémaalapú beszédfelismerő rendszereket tanítunk magyar nyelvű híradós adatbázison. A felismerőnk akusztikus modellként egy modern mély neuronháló struktúrát alkalmaz, nyelvi modell oldalon pedig a hagyományos n-gram megközelítést hasonlítjuk össze mély rekurrens hálókkal. Eredményeink alapján kijelenthetjük, hogy a morfémaalapú nyelvi modell használatával nem csak a szótár méretét csökkentettük, de a felismerés pontosságot is szignifikánsan javítottuk.

2. Kapcsolódó irodalom

Morfémaalapú rendszer esetén első lépésként szegmentálnunk (a szavakat morfémákra bontani) kell a tanítóadatunkat, ezt többféle módon is megtehetjük. A szegmentáláshoz használhatunk nyelvspecifikus szabályokon és szótáron alapuló módszert, például a HunMorph (Trón és mtsai, 2005) rendszer alkalmazásával.

Alternatívaként használhatunk statisztikai szegmentáló eljárást is, ennek előnye, hogy nem igényel semmilyen külső tudást, a rendelkezésére álló szöveget felhasználva keres egy optimális felbontást. Ezen módszerek közül mi a Morfessor Baseline (Creutz és Lagus, 2002) eljárást használtuk, amely egy Minimum Description Length (MDL) elven működő módszer. Célja, hogy felügyelet nélkül létrehozzon egy optimális lexikont, amely segítségével szegmentálható a tanító szöveg.

Magyar nyelvű beszédfelismerésen belül morfémaalapú nyelvi modell használatával már több mű is foglalkozott (Mihajlik és mtsai, 2007; Németh és mtsai, 2007; Tarján és mtsai, 2009; Tarján és mtsai, 2014), melyek több lehetséges szegmentálási módszert hasonlítanak össze. Eredményeikből megállapítható, hogy a Morfessor Baseline módszer képes hatékonyan szegmentálni magyar nyelvű szövegeket. Az eddigi munkákban közös, hogy nyelvi modellként a hagyományos n-gram módszert alkalmazták, ezzel ellentétben mi mély rekurrens neuronhálókat is alkalmaztunk kísérleteink során.

A közelmúltban megmutatták, hogy más nyelveken (finn és észt) is számottevő javulások érhetőek el automatikusan konstruált morféma szintű nyelvi modell használatával (Smit és mtsai, 2017). A javasolt eljárásukban a Morfessor Baseline-t alkalmazták a szegmentálási lépés során, majd n-gram modelleket hasonlítottak össze rekurrens neuronhálókkal, vizsgálataink során mi is ezt a módszert követtük.

3. Morfémák szegmentálása

Szavak szegmentálása során célunk meghatározni, hogy az egyes szak mely morfémákból épülnek fel. A feladat elvégzésére alkalmazhatunk nyelvspecifikus szabályalapú rendszereket vagy automatikus módszereket, esetleg ezek kombinációját. Fontos megjegyezni, hogy mi az automatikus módszerekre fókuszáltunk, az általuk javasolt egységek azonban nyelvészeti szempontból nem feltétlenül tekinthetőek morfémáknak, de az egyszerűség kedvéért mi morfémaként fogunk ezekre az egységekre hivatkozni.

Az itt alkalmazott Morfessor Baseline algoritmus a felügyelet nélküli módszerek családjába tartozik. Tanítás során egy mohó, lokális keresést hajt végre az optimális morféma lexikon meghatározásához, amely a következő hibafüggvény optimalizálja:

$$L(\Theta, D_w) = -logp(\Theta) - \alpha logp(D_w|\Theta), \tag{1}$$

ahol Θ a modell paraméterei, D_w a tanító adat, α pedig a hibafüggvény paramétere. A prior valószínűség ($p(\Theta)$) kizárólag a lexikontól függ, számítása MDL alapú módszerrel történik (Virpioja és mtsai, 2013). Az adat likelihood valószínűségét a tanító adatbázisban található szavak aktuális analízise ($Y = (y_1 \dots y_N)$) alapján becsülhetjük;

$$p(D_w|\Theta) = \sum_{j=1}^{N} logp(w_b) \sum_{i=1}^{|y_j|} logp(m_{ji}|\Theta), \tag{2}$$

ahol m_{ji} a j-edik szó felbontásának i-edik morfémája, w_b pedig a szavak közötti határoló szimbólum. Az α paraméter segítségével tudjuk kontrollálni a lexikonban található morfémák számát, kicsi érték esetén a prior lesz a meghatározó tag, így az optimalizáló próbál minél kisebb lexikont létrehozni. Nagy α érték esetén a likelihood lesz a domináns, ami miatt a modell hosszú morfémákat preferál, ez pedig nagyobb lexikont eredményez.

A tanítás kezdetén az összes szó, amely előfordul a tanító adatbázisban bekerül a lexikonba, majd az algoritmus kiválaszt ezek közül egyet, amelynek megkeresi az optimális felbontását a 1. képlet alapján. Az algoritmus ez után iteratívan folytatja a felbontások keresését, amíg egy optimális lexikont nem kap.

A tanítási lépés után a dekódolási lépés következik, amikor is szavakat próbálunk morfémákra bontani, a legvalószínűbb felbontás meghatározására a Viterbi algoritmust használhatjuk.

Kísérleteink során a Morfessor-2.0 (Virpioja és mtsai, 2013) szoftvert használtuk a szegmentáló modell létrehozására. Az egyszerűség kedvéért csak a szegmentálás végrehajtása után, a nyelvi modell tanítás során különböztettük meg a prefix, szuffix és közbülső morfémákat. A 1. táblázat egy példa mondat szegmentálását tartalmazza. Megfigyelhető, hogy az α értékének csökkenésével egyre kisebb egységekre bontja a modell a szavakat.

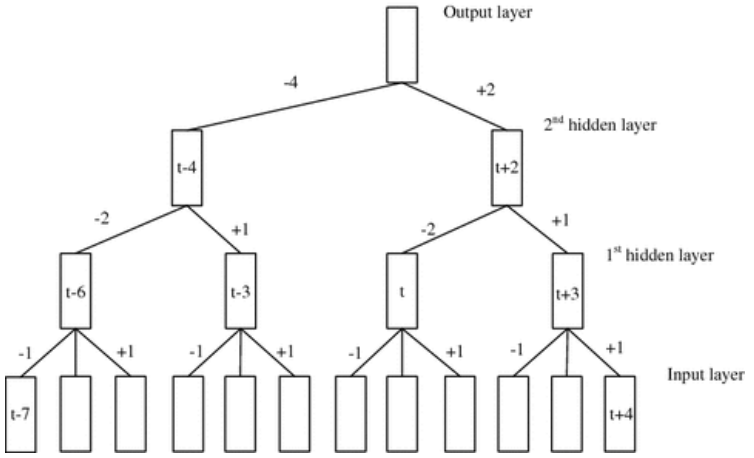
α	szegmentált példamondat
0.1	közösség+ +ét minden oldalról fenyegető veszélyeket
0.01	közösség+ +ét minden oldalról fenyegető veszély+ +eket
0.001	közös+ +ség+ +ét minden oldal+ +ról fenyeget+ +ő veszély+ +eket

1. táblázat. Példa szegmentálásra különböző α paraméterek esetén.

4. Akusztikus modell

Egy standard akusztikus modell feladata, hogy a bementi spektrális jellemzők alapján megbecsülje az egyes fonémák valószínűségét. Tanítás során a kiejtési szótár segítségével határozzuk meg az egyes szakhoz tartozó fonémákat, ez a megközelítés sajnos esetünkben nem alkalmazható, mivel a nyelvi modellünk morféma szinten működik. A problémát az okozza, hogy minden morfémához definiálnunk kellene annak kiejtését a kontextus (környező morfémák) ismerete nélkül. Szerencsére a probléma viszonylag könnyen kezelhető, amennyiben fonémák helyett grafémákat használunk akusztikus egységként, ebben az esetben a kiejtési szótár könnyen generálható.

Kísérleteinkben graféma alapú akusztikus modelleket használtunk, amelyeket a Kaldi (Povey és mtsai, 2011) rendszer segítségével tanítottunk. Végső modellként egy időkésleltett neuronhálót (time-delay neural network, TDNN) (Peddinti és mtsai, 2015) használtunk, amelyet lattice-mentes maximális kölcsönös információ (lattice-free maximum mutual information)(Povey és mtsai, 2016) módszerrel tanítottunk.



1. ábra: Egy három réteges TDNN neuronháló struktúrája.

A TDNN hálók specialitása, hogy rejtett rétegeik időbeli konvolúciót végeznek, az első rejtett réteg csak egy kis időbeli kontextust dolgoz fel, a későbbi rétegek pedig egyre nagyobb időablakot fednek le a korábbi rejtett rétegek segítségével. Működését a 4. ábra szemlélteti. Tanításuk során a Kaldi keretrendszerben elérhető ún. chain receptet követtük. A neuronháló 10 rejtett réteget tartalmazott, amelyek mindegyike 1000 darab relu aktivációs függvényt alkalmazó neuronból állt. Bemenetként standard MFCC jellemzővektorokat használtunk, összesen 13 koefficienst illetve azok Δ-ját és $\Delta\Delta$-ját.

5. Nyelvi modell

Tradicionálisan nyelvi modellezésre az ú.n. n-gram modelleket szokás használni, amelyek az előző n – 1 darab szó alapján becsülik a következő szó valószínűségét. Ezen modellek tanítása során a szükséges statisztikákat a rendelkezésre álló szövegből számítjuk. A pontosabb eredmények elérése érdekében több finomítása is létezik a módszernek, mi ezek közül a Kneser-Ney simítást alkalmaztuk a VariKN (Siivola és mtsai, 2007) rendszer használatával. Kísérleteink során a hagyományos 3-gram modellek mellett számottevően nagyobb n-gram-okat is felhasználtunk, abban bízva, hogy morfémaalapú modellek esetén hasznos lehet a nagyobb kontextus használata.

A hagyományos n-gram megközelítés mellett a manapság nagy népszerűségnek örvendő rekurrens neuronhálókat is kipróbáltuk. Az utóbbi években a rekurrens neuronhálók kiemelkedően jó eredményeket értek el természetes nyelvi feldolgozásban. Beszédfelismerésben a rövid- és hosszú-távú memória cellákat (long short-term memory, LSTM) alkalmazó változatuk terjedt el leginkább(Young és mtsai, 2018). A legfőbb különbség a hagyományos rekurrens neuron és az LSTM cella között, hogy utóbbi nem csak a korábbi kimenetét kapja meg bemenetként, hanem rendelkezik egy belső állapottal is, amely a hosszú-távú emlékezésben segít.

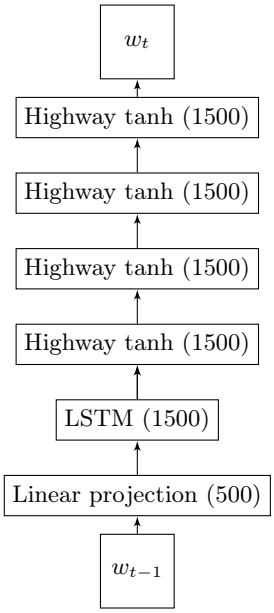
Formálisan, egy bemeneti vektor (x_{t-1}) esetén egy LSTM cella első lépésben a következő számításokat végzi:

$$\begin{aligned} f_t &= \sigma(W_f x_{t-1} + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_{t-1} + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_{t-1} + U_o h_{t-1} + b_o), \end{aligned} \tag{3}$$

ahol h_{t-1} az előző kimenet, σ pedig a sigmoid függvény. A kiszámított bemeneti (i_t), kimeneti (o_t) és felejtő (f_t) kapuk értékei alapján pedig a végső kimenet (h_t) illetve a belső memória (c_t) új értéke kerül meghatározásra;

$$\begin{aligned} c_t &= f_t c_{t-1} + i_t tanh(W_c x_{t-1} + U_c h_{t-1} + b_c) \\ h_t &= o_t tanh(c_t) \end{aligned} \tag{4}$$

Munkánkban a nyelvi modellként használt neuronhálók struktúráját a 2 ábra szemlélteti. Első lépésben a bementet egy projekciós réteg dolgozza fel, amely a beágyazást (embedding) végzi, ezt a réteget nem tanítottuk külön, a tanítás elején véletlenszerűen inicializáltuk. A beágyazó réteg után következik az LSTM réteg, ami a belső memória segítségével próbál információt tárolni a korábbi szavakról vagy morfémákról, majd négy highway réteg dolgozza fel ennek kimenetét. A highway rétegek lényege, hogy kimenetük az eredeti bemenet és a rejtett neuronok kimenetének lineáris kombinációja, ez megkönnyíti a gradiens propagálását tanítás során, ami pedig lehetővé teszi, hogy sok rejtett réteget használjuk hatékonyan. A lehetséges következő szavak valószínűségeit egy softmax réteg segítségével becsüljük, a neuronhálók tanításhoz a TheanoLM (Enarvi és Kurimo, 2016) keretrendszert használtuk.



2. ábra: A kísérleteink során használt rekurrens nyelvi modell felépítése.

5.1. Kiértékelés neuronhálós nyelvi modellel

A felismerési folyamat során sajnos nem realisztikus egyből a neronhálós nyelvi modellt használni, hiszen ismert, hogy a dekódolás keresési tere exponenciálisan növekszik a hipotézis hosszával, ez pedig lelassítja a rendszert. További ellenérv, hogy a neuronhálók kiértékelése számottevően több időt igényel mint egy egyszerűbb n-gram használata. Ezen problémára több megoldás is létezik, az egy lehetőség, hogy a neuronháló felhasználásával szöveget generálunk, melyből hagyományos n-gram modellt tanítunk és ezt használjuk a felismerés során (Mittul és mtsai, 2018; Tarján és mtsai, 2019), így ugyan veszítünk némi információt, de lehetőségünk van gyors, akár online dekódolásra is.

Talán a legelterjedtebb megoldás mégis a kétkörös dekódolás (two pass decoding). Ekkor első körben egy egyszerű n-gram nyelvi modell (tipikusan 3-gram) segítségével ún. lattice-t hozunk létre, majd a második körben újrasúlyozzuk (rescore) a felismerési hipotéziseket a lattice-ben a neuronháló kimenetei alapján. Kísérleteinkben mi is ezt a megközelítést alkalmaztuk, hiszen így tisztább képet kaphatunk a neuronhálók pontosságáról.

Alternatívaként használhatunk n-legjobb listákat (n-best list) (Deoras és mtsai, 2011), azonban kezdeti kísérleteink alapján ez a megközelítés rosszabb eredményeket ad mint a kétkörös módszer. Megemlítenénk, hogy közelmúltban megjelentek új módszerek, amelyek képesek a dekódolást csak neuronhálós nyelvi modellel hatékonyan végrehajtani (Jorge és mtsai, 2019), sajnos ezt a megközelítést nem volt időnk tesztelni.

Nyelvi modell egysége	Szótár méret teszt	OOV ráta
Szó	420520	9.9%
Morf. α=0.1	183803	0.5%
Morf. α=0.01	53667	0.3%
Morf. α=0.001	11562	0.2%

2. táblázat. Tanító adatbázis statisztikái.

6. Tanító adatbázisok

Az akusztikus modellek tanítására az Origo korpuszt használtuk, amely összesen 2.7 millió mondatot tartalmaz, a szóalakok száma pedig meghaladja az 50 milliót. A Morfessor modellek tanítása előtt véletlenszerűen kiválasztottunk 10000 mondatot, ezeket validációs halmazként használtuk.

Az akusztikus modell tanításához egy magyar nyelvű híradós adatbázist (Tóth és Grósz, 2013) használtunk, amely megközelítőleg 30 órányi beszédanyagot tartalmazott, ebből 2 órányit használtunk validációs, 4 órányit pedig teszt halmazként.

7. Eredmények

Első lépésben a szószintű és a morféma szegmentálással kapott szótárakat hasonlítottuk össze (2. táblázat). Ezek létrehozása során kizárólag a szöveges adatbázist használtuk (az akusztikus tanítóadat átirata nem lett hozzáadva a tanítóadathoz). A szószintű megközelítés esetén a VariKN rendszert használtuk a szótár létrehozására, a kiválasztott nagyjából 420000 szavas szótár a szöveges tanítóadat leggyakoribb szavaiból lett kiválasztva, ez az akusztikus teszthalmazban található szavak 9.9%-át nem tartalmazza. Természetesen nagyobb szótár esetén ez az arány csökkenthető, ám ekkor a nyelvi modell mérete számottevően megugrik, különösen a nagy n-gram esetén.

Morfémaalapú megközelítések esetén látható, hogy sokkal kisebb szótárral is sokkal jobban le tudjuk fedni a teszt adatot, ezzel lehetővé téve a pontosabb felismerést. Ahogy egyre jobban csökken a lexikon mérete (annak eredményeként, hogy a prior tagra koncentrál a szegmentáló algoritmus), egyre kevesebb szót találunk a teszt halmazban, amit nem tudunk a morfémákkal lefedni (out-of-vocabulary, OOV arány). Természetesen a kisebb szótár azt is jelenti, hogy egyre kisebb egységekre bontjuk az egyes szavakat, ami nem feltétlenül előnyös a nyelvi modell számára.

Vizsgálataink során három különböző nyelvi modellt alkalmaztunk, a felismerés első fázisát mindig a 3-gram modellel végeztük. A második körben pedig egy nagy n-gram modellt illetve a neuronhálós rendszerünket használtuk. Az összehasonlításokhoz a szóhiba-arány (word error rate, WER) metrikát használtuk, a morféma alapú felismerő kiértékelésénél a WER ugyanazt a szószintet jelenti-e, mint a szóalapúnál. A szóalakok rekonstrukciójához a felismerés végén a morfémákat a '+' határoló jelzés esetén összevontuk.

Nyelvi modell egysége	Nyelvi modell típusa	Validációs halmaz	Teszt
Szószintű	VariKN (3-gram)	20.91%	19.73%
	VariKN (16-gram)	20.95%	19.65%
	LSTM	19.30%	17.98%
Morfessor α=0.1	VariKN (3-gram)	17.60%	16.17%
	VariKN (16-gram)	17.48%	16.17%
	LSTM	16.69%	15.29%
Morfessor α=0.01	VariKN (3-gram)	18.92%	17.48%
	VariKN (21-gram)	19.00%	17.49%
	LSTM	15.28%	14.09%
Morfessor α=0.001	VariKN (3-gram)	19.18%	18.00%
	VariKN (24-gram)	18.70%	17.44%
	LSTM	15.69%	14.41%

3. táblázat. Beszédfelismerési eredmények.

A 3. táblázatban láthatóak a különböző megközelítésekkel elért eredményeink. A szószintű rendszereket tekintve megállapítható, hogy nagy méretű (16-gram) modell használata nem javít a felismerés pontosságán, a neuronhálós megoldás viszont szignifikánsan jobb eredményt képes produkálni, mint amit n-gram használatával el tudunk érni. Ez utóbbi megfigyelés a morfémaalapú rendszerek esetén is igaz. Morfémákat alkalmazó felismerők minden esetben jobban teljesítettek mint a hagyományos szószintűek, így megállapíthatjuk, hogy magyar nyelvű beszéd esetén célszerű használatuk.

Érdekességként megfigyelhető, hogy kicsi α esetén, amikor is a szavakat sok kicsi egységre bontjuk, akkor a 23-gram modell már jobban teljesít mint a sima 3-gram. Ennek magyarázata abban keresendő, hogy ekkor már fontos a nagy kontextus használata, hiszen a 3-gram használatával előfordulhat, hogy hosszabb szavakat (amik több mint 3 morfémára lettek bontva) nem tudunk lefedni és így semmi információval nem rendelkezünk a korábbi szavakról.

A legjobb eredményeket neuronhálós nyelvi modellel értük el $\alpha = 0.01$ használatával. Ekkor 3.9% javulást láthatunk a szószintű változathoz hasonlítva, ami közel 22%-os relatív javulást jelent. A magyarázat arra, hogy miért pont ez a szegmentálás bizonyult legjobbnak az lehet, hogy ekkor már kellően lecsökkent a szótár mérete ahhoz, hogy hatékonyan tudjon a neuronháló tanulni és a szavakat nem bontottuk túl sok egységre, így nem jelentet túl nagy kihívást a korábbi morfémákra való "emlékezés" sem.

Megfigyelhető továbbá, hogy egyre kisebb morfémaszótár esetén az n-gramok egyre rosszabb eredményt értek el. Ebből arra lehet következtetni, hogy ezen modellek a nagy méretű morfémákat preferálják, ami nagy szótárat eredményez.

8. Konklúzió

Cikkünkben morfémaalapú rekurrens nyelvi modelleket alkalmazó beszédfelismerők teljesítményét vizsgáltunk egy magyar nyelvű korpuszon. Megállapítható,

hogy a szavak felbontása morfémákra megkönnyíti a nyelvi modell feladatát, így pontosabb felismerő rendszereket taníthatunk. A morfémákat alkalmazó modellek előnye a szószintűekkel szemben két fő tényezőnek köszönhető, egyrészről a lényegesen kisebb felismerési szótárnak, másrészről pedig annak, hogy morfémák segítségével lényegesen több szót tudunk felépíteni így csökkentve az OOV rátát. Fontos azonban megtalálni az egyensúlyt a szótár és a morfémák mérete között, hiszen a túl kicsi egységekre bontás ugyan lényegesen csökkenti a lexikon méretét, de nehezebbé is teszi a pontos modell tanítását.

Eredményeink alapján az is nyilvánvaló, hogy a hagyományos n-gram modelleknél számottevően jobban teljesítenek a neuronhálót alkalmazók, ahogy ezt már több korábbi munka is igazolta. További kutatásaink során a neuronhálós nyelvi modell továbbfejlesztésére tervezünk fókuszálni. Érdekes kérdés például, hogy vajon a szószintű modellek esetén rendkívül jól teljesítő figyelem (attention) mechanizmus (Bahdanau és mtsai, 2015) vajon morfémaalapú rendszer esetén is hasznos-e?

Hivatkozások

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.0473

Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6. pp. 21–30. MPL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), https://doi.org/10.3115/1118647.1118650

Deoras, A., Mikolov, T., Church, K.: A fast re-scoring strategy to capture long-distance dependencies. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1116–1127. Association for Computational Linguistics, Edinburgh, Scotland, UK. (Jul 2011), https://www.aclweb.org/anthology/D11-1103

Enarvi, S., Kurimo, M.: TheanoLM — An Extensible Toolkit for Neural Network Language Modeling. In: Interspeech 2016. pp. 3052–3056 (2016), http://dx.doi.org/10.21437/Interspeech.2016-618

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., és mtsai: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29(6), 82–97 (2012)

Jorge, J., Giménez, A., Iranzo-Sánchez, J., Civera, J., Sanchis, A., Juan, A.: Real-Time One-Pass Decoder for Speech Recognition Using LSTM Language Models. In: Proc. Interspeech 2019. pp. 3820–3824 (2019)

Mihajlik, P., Fegyó, T., Tüske, Z., Ircing, P.: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages - like Hungarian. In: Interspeech 2007. pp. 1497–1500 (2007)

Mittul, S., Peter, S., Sami, V., Mikko, K.: First-pass decoding with n-gram approximation of RNNLM: The problem of rare words. In: Machine Learning in Speech and Language Processing Workshop (2018)

Németh, B., Mihajlik, P., Tikk, D., Trón, V.: Statisztikai és szabály alapú morfológiai elemzők kombinációja beszédfelismerő alkalmazáshoz. In: Magyar Számítógépes Nyelvészeti Konferencia. pp. 95–105 (2007)

Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: INTERSPEECH (2015)

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (Dec 2011), iEEE Catalog No.: CFP11SRW-USB

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S.: Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In: INTERSPEECH (2016)

Siivola, V., Creutz, M., Kurimo, M.: Morfessor and VariKN machine learning tools for speech and language technology. In: INTERSPEECH. pp. 1549–1552. ISCA (2007)

Smit, P., Virpioja, S., Kurimo, M.: Improved subword modeling for wfst-based speech recognition. In: Proc. Interspeech 2017. pp. 2551–2555 (2017)

Tarján, B., Fegyó, T., Mihajlik, P.: A bilingual study on the prediction of morph-based improvement. In: Spoken Language Technologies for Under-Resourced Languages (2014)

Tarján, B., Fegyó, T., Mihajlik, P.: Ügyfélszolgálati beszélgetések nyelvmodellezéserekurrens neurális hálózatokkal. In: Magyar Számítógépes Nyelvészeti Konferencia. pp. 23–33 (2019)

Tarján, B., Mihajlik, P., Tüske, Z.: Nagyszótáras híranyagok felismerési pontosságának növelése morfémaalapú, folyamatos beszédfelismerővel. In: Magyar Számítógépes Nyelvészeti Konferencia. pp. 185–194 (2009)

Tóth, L., Grósz, T.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: Text, Speech, and Dialogue. pp. 36–43. Springer Berlin Heidelberg (2013)

Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: Open source word analysis. In: Proceedings of Workshop on Software. pp. 77–85. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://www.aclweb.org/anthology/W05-1106

Virpioja, S., Smit, P., Grönroos, S.A., Kurimo, M.: Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. D4 julkaistu kehittämistai tutkimusraportti tai -selvitys (2013), http://urn.fi/URN:ISBN:978-952-60-5501-5

Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine 13(3), 55–75 (Aug 2018)

A depresszió hang alapú felismerésének optimalizációja hangfelvétel hossz alapján

Pašić Azra[1], Kiss Gábor[2], Sztahó Dávid[2]

[1]Karlsruhe Institute of Technology
[2]Budapest Műszaki és Gazdaságtudományi Egyetem
uwhvy@student.kit.edu, {kiss.gabor, sztaho}@tmit.bme.hu

Kivonat A depresszió komoly hangulatzavar, amely világszerte már a lakosság több mint 3%-át érinti, és ez a szám feltehetően tovább fog nőni az elkövetkezendő években, évtizedekben. A depresszió diagnosztizálása maga is egy komoly feladat, amely jelenleg kizárólag a terület szakembereire hárul, akikből pedig egész bizonyosan nincs elég. Ebben a helyzetben nagy jelentőséggel bírhat egy olyan automatizált depresszió felismerési rendszer bevezetése, amely nagymértékben asszisztálni tudná a szakemberek munkáját a diagnosztizálás során. E cikkben bemutatunk egy, a depresszió osztályozására fejlesztett hang-alapú felismerő rendszert, amely ötvözi az akusztikai jellemzők kinyerését, a jellemzőkiválasztást és a szupport vektor gépek hiperparaméter-optimalizációját. Természetesen, a hang-alapú modellhez szükséges egy optimális hangfelvétel hossz meghatározása is, mely kompromisszumot jelent a felismerőrendszer igényei és a páciensek kényelme között. A modell hatékonyságát különböző hosszúságú felvételeken vizsgáltuk, hogy belátást nyerjünk abba, hogy a felvétel-hossz miként és milyen mértékben befolyásolja a felismerés pontosságát.
Kulcsszavak: depresszió, beszédjel alapú detektálás, osztályozás, szupport vektor klasszifikáció

1. Bevezetés

A súlyos depresszív zavar (legtöbbször csak „depresszió”) olyan mentális zavar, amely a levertség, reménytelenség, szorongás és kitartó szomorúság tüneteivel jár (Association et al., 2013) (Cummins et al., 2015). Világszerte már a lakosság több mint 3%-át érinti (Andrade et al., 2003), és ez a szám feltehetően tovább fog nőni az elkövetkezendő években, évtizedekben. A betegség hatása az érintettek életminőségére olyan krónikus megbetegedésekhez lett hasonlítva mint a cukorbetegség és a magas vérnyomás (Hays et al., 1995). Ezen kívül pedig a depressziós betegeknél húszszor nagyobb az esély az öngyilkosságra mint az egészséges lakosságnál (Lépine and Briley, 2011). Mindezek ellenére a depresszió nagyon is kezelhető betegségnek számít, de ehhez szükséges az időszerű felismerés. Gyógyulás után is érdemes a korábbi betegekkel foglalkozni, mivel a visszaesés veszélye nagy, és az első depressziós epizódtól szenvedők 80%-a legalább még egyet tapasztal élete folyamán (Lépine and Briley, 2011).

Mivel a depresszió diagnosztizálása és szűrése is szakemberekhez van kötve, folytonos a pszichológus és pszichiáter hiány, ami ahhoz is vezet, hogy a depressziós betegek nagy része nem is kerül felismerésre (Lépine and Briley, 2011) — annak ellenére, hogy a kezelés elmaradása megötszöröz az öngyilkosság esélyét (Strakowski and Nelson, 2015). Ebben a helyzetben nagy jelentőséggel bírhat egy olyan automatizált depresszió felismerési rendszer bevezetése, amely nagymértékben asszisztálni tudná a szakemberek munkáját a diagnosztizálás során. A diagnosztikai eljárásban az orvos megfigyeli a betegnek a kinézetén, a viselkedésén és a hangulatán kívül a beszédét — ezen belül pedig a hangját, hangzását is (Association et al., 2013). Ebből kifolyólag a depresszió automatikus felismerése hang alapján sokat ígérő ötlet. A depresszió és a beszéd kapcsolata már az 1980-as évektől kutatott, és több akusztikai illetve fonetikai paramétert kapcsolatba hoztak a depresszióval (Nilsonne, 1988).

E cikkben bemutatunk egy, a depresszió osztályozására fejlesztett hang-alapú felismerő rendszert, amely ötvözi az akusztikai jellemzők kinyerését, a jellemzőkiválasztást és a szupport vektor gépek hiperparaméter-optimalizációját. Természetesen, a hang-alapú modellhez szükséges egy optimális hangfelvétel hossz meghatározása is, mely kompromisszumot jelent a felismerő-rendszer igényei és a páciensek kényelme között (Rutowski et al., 2019). A modell hatékonyságát különböző hosszúságú felvételeken vizsgáltuk, hogy belátást nyerjünk abba, hogy a felvétel-hossz miként és milyen mértékben befolyásolja a felismerés pontosságát. Gépi tanulással kétféle felismerés valósítható meg: az osztályozás, amely a depressziós állapotot becsüli meg, és a regresszió, amely annak a súlyosságáról kísérel információt adni. Ebben a cikkben az osztályozást használtuk, melyet szupport vektor gépekkel valósítottunk meg.

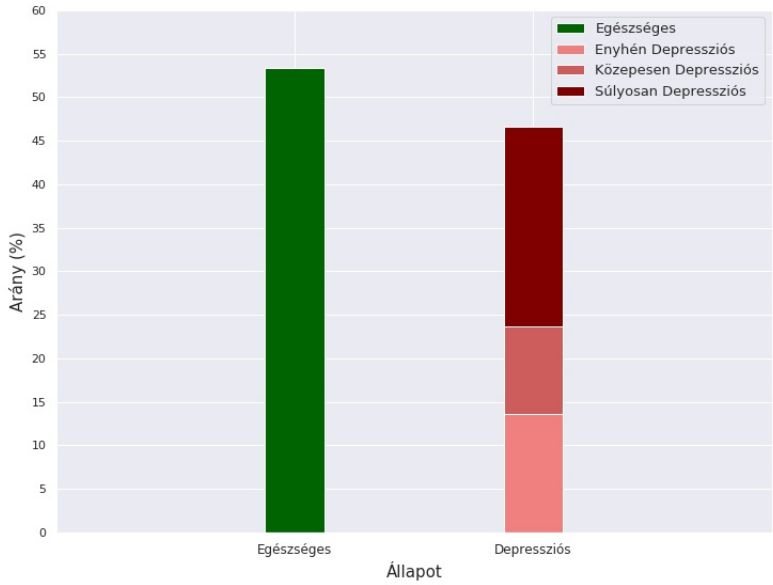
A cikk bevezetés utáni felépítése a következő: először bemutatjuk a beszéd adatbázist amivel dolgoztunk, majd a kutatásban felhasznált módszereinket – ezen belül az előfeldolgozást, a jellemző kinyerést, az osztályozást és a tesztelést is. Ezután következik az eredmények bemutatása és tárgylása, valamint az összegzés és a konklúzió.

2. Adatbázis

A beszédminták gyűjtése a Semmelweis Egyetem Pszichiátriai és Pszichoterápiás Klinikájával együtt lett végezve. A beszélők lefedik a depresszió súlyosságának különböző fokozatait, az egészséges állapottól az egészen súlyos depresszióig. A vizsgált személyek egy fonetikusan kiegyensúlyozott mesét („Az északi szél és a nap") olvastak fel, amely széles körben elterjedt a hasonló vizsgálatokban. A felvételek csendes helyiségben kerültek rögzítésre, 44.1 kHz mintavételi frekvenciával. Az adatbázisba gyűjtött felvételekhez el lett készítve a fonéma szintű szegmentálás, a labor által fejlesztett automatikus szegmentáló program segítségével (Kiss et al., 2013).

A depresszió súlyossága is minden esetben rögzítésre került – a két legelterjedtebb skála a Hamilton Rating Scale for Depression (HAMD) (Williams, 1988) és a Beck Depression Index (BDI) (Beck et al., 1996). Mi a BDI továbbfejlesztett

változatát használtuk, a BDI-II skálát (Beck et al., 1996). A BDI-II skála pontszámaihoz a következő besorolás adott: 0-13 egészséges, 14-19 enyhe depresszió, 20-28 közepes depresszió, 29-63 súlyos depresszió. A BDI pontszámok 0-tól 50-ig fordultak elő az adatbázisban. Az adatbázis 118 hangfelvételt tartalmazott, ebből 55 depressziós és 63 egészséges mintát. A különböző súlyosságok előfordulása az 1. ábrában adott. A vizsgált személyek átlagéletkora 42,5 év (min.: 20; max.: 70; std: 14,5).



1. ábra: Az egészséges és depressziós minták eloszlása az adatbázisban

3. Módszerek

A jellemző-kinyerés Python 2.7 programmal lett végezve (Python, 2007). A librosa és soundfile csomagok a felvételek kezeléséhez és az akusztikai jellemzők kinyeréséhez lettek felhasználva (McFee et al., 2015). További jellemzők a parselmouth (Jadoul et al., 2018) csomaggal kerültek kinyerésre, amely a Praat program C++ kódjából kinyert Python változata (Boersma et al., 2002). A parselmouth-tal együtt lett használva a tgt csomag, amely a Praat által generált Textgrid fájlok (ezek tartalmazzák a szegmentálást) kezeléséhez volt szükséges (Buschmeier and Wlodarczak, 2013). A különböző klasszifikációs modellek a

LibSVM könyvtárral lettek felépítve (Chang and Lin, 2011). A hiperparaméter optimalizáció Grid Search algoritmussal lett végezve, amely a lehető paraméterkombinációkból a legjobbat választja ki.

3.1. Előfeldolgozás

A felvételek először 16 kHz-en újra lettek mintavételezve. A BDI-II pontszámuk alapján a minták a depressziós és egészséges csoportokba lettek sorolva és az alapján felcímkézve. Ezt követően a szegmentálás segítségével a felvételek három részre lettek osztva, majd ezekből lett képezve a három vizsgált hossz – az egy harmad, két harmad és egész felvétel – méghozzá úgy, hogy csak mondat végén történtek a vágások. Ez azért volt lényeges, mert az idő alapú szeparáció amely nem veszi figyelembe a mondathatárokat torzította volna az akusztikai jellemzőket. Továbbá ez azt is jelenti, hogy az egy harmad és két harmad nem szó szerint értendő (az egy harmad felvétel valamivel rövidebb, mint a két harmad felvétel fele). A hasonló kutatásokban használatos jellemzők alapján ezek a paraméterek kerültek kiszámításra a felvételeken (Kiss and Vicsi, 2017) (Kiss and Vicsi, 2014) (Cummins et al., 2015) (Alghowinem et al., 2013): formáns frekvenciák (F1, F2, F3), mel-skálás spektrogram, mel-frekvenciás kepsztrális együtthatók (MFCC-k, 10 koefficienssel), chromagram, tonal centroid, valamint különböző intenzitás, frekvencia és hangmagasság értékek a Praat-ból (jitter, shimmer, number of voice breaks, fraction of locally unvoiced frames, degree of voice breaks). A jellemzők -1 és 1 közötti értékekre lettek normalizálva.

3.2. Jellemzők kiválasztása

Az algoritmusok pontosságát nagyban befolyásolja a megfelelő jellemzők kiválasztása, vagyis a lényegtelen jellemzők elhagyása. Ez főleg fontos kis mintahalmaz esetén, mint amilyen a miénk is. Az optimális jellemzők Fast Forward Selection-nel kerültek kiválasztásra. Az eljárás során az i-dik lépésben rendelkezésre áll az algoritmus szerint optimális i-1 hosszú jellemzővektor, amihez ezután egyesével hozzá lesznek adva a még fel nem használt jellemzők és k-fold kereszt validáció alapján (default hiperparaméterekkel) az i hosszúságú jellemzővektor közül ki lesz választva az, amely a legnagyobb pontosságot adta (Mao, 2002). Az eljárás hátránya, hogy ha egy lépésben egy jellemző be lesz választva a jellemzőhalmazba, az minden halmazban benne lesz, viszont a jellemző kiválasztás gyors (Mao, 2002).

3.3. Osztályozás

A szupport vektor gépek alapelve, hogy a címkézett példákat (azaz a training készletet) térbeli pontokként jelenítse meg, oly módon, hogy az osztályok a lehető legjobban el legyenek különítve (Cortes and Vapnik, 1995). Ezt követően az új adatpontokat ugyanabba a térbe térképezi fel, és attól függően, hogy az osztályok közötti rés melyik oldalára esnek, a két kategória egyikébe lesznek sorolva (Cortes and Vapnik, 1995). Lineárisan nem szeparálható problémák esetén kernel

függvény segítségével a probléma nagyobb dimenzitású térbe kerül, amelyben szeparálhatóvá alakul (Cortes and Vapnik, 1995). Különböző kernel függvények léteznek, mint például a polinomiális, a szigmoid és a radiális (Cortes and Vapnik, 1995). A kutatás során c-SVC algoritmust radiális (Radial Basis Function) kernellel használtunk, különböző gamma együtthatókkal és C értékekkel (a C határozza meg az osztályok minél nagyobb elkülönülésének és a hibás oldalra eső minták számának a trade-off-ját). Ezek a hiperparaméterek Grid Search algoritmussal lettek kiválasztva, amely kipróbál minden kombinációt és kiválasztja a legjobban teljesítő hiperparaméter-párt.

3.4. A tesztelési eljárás

Az adatbázis alacsony mintaszáma miatt az ebben az esetben szokásos k-fold keresztvalidáció (k-Fold Cross Validation) (Kohavi et al., 1995) lett használva a tesztelések során (mint ahogy az FFS és a Grid Search során is). A keresztvalidációs eljárás a mintahalmazt k egyenlő részre osztja, majd mindegyik csoportot egyszer teszthalmazként használ, a megmaradó részeket $(k-1)$ pedig tanítóhalmazként. A teszthalmazokon kapott eredmények átlaga jellemzi az egész rendszer pontosságát. A modell jellemzésére tévesztési mátrixokat is bemutatunk, amelyekből kivehető, hogy az egészséges és a depressziós mintákat külön-külön mennyire jól ismeri fel a modellünk.

4. Eredmények

A kísérleteket egy harmad, két harmad és egész felvételeken végeztük, a jellemzőkinyerés és normalizálás után a jellemzővektorokat Fast Forward Selection-nel kaptuk meg, majd ezeken tanítva a Grid Search-et megtaláltuk az optimális hiperparamétereket a szupport vektor osztályozáshoz. A tesztelési eljárás során minden esetben 10 részre osztottuk az adathalmazt, és a teljesítmény értékeléséhez a pontosságot (a helyesen osztályozott minták számának és az összes minta számának hányadosát) használtuk. A következő táblázatban láthatóak a különböző hosszúságú felvételeken elért pontosságok és a hibásan osztályozott minták számának relatív csökkenése (az egy harmad felvételhez képest).

	Egy harmad	Két harmad	Egész felvétel
Pontosság	88%	90%	92%
Hiba relatív csökkenése	-	17%	33%

1. táblázat. Az elért pontosságok és a hiba relatív csökkenése

Egy harmad felvételen a legjobb paramétereknek bizonyultak a C=1 és g = 0.125. A tíz kiválasztott jellemző között voltak koefficiensek az MFCC-ből, a chromagramból, a mel-skálás spektrogramból, a contrastból, valamint a shimmer,

a number of voice breaks (egymást követő impulzusok közötti hosszabb szünetek száma) és a formáns frekvenciák is. A következő táblázatban láthatóak az egy harmad felvételen elért eredmények tévesztési mátrix formájában.

	Osztályozott egészséges	Osztályozott depressziós
Tényleges egészséges	92.1%	7.9%
Tényleges depressziós	16.4 %	83.6%

2. táblázat. Az egy harmad felvételen kapott tévesztési mátrix

A két harmad felvételen számított hiperparaméterek kevéssel eltérnek az előbbitől: C=2, g = 0.25. A jellemzőknél azonban nagy a hasonlóság – továbbra is a tíz kiválasztott között volt az MFCC, a chromagram, a mel-skálás spektrogram és a number of voice breaks, de ebben az esetben beválasztásra került egy koefficiens a tonal centroid-ból is. A 3-as számú táblázatban láthatóak az eredmények.

	Osztályozott egészséges	Osztályozott depressziós
Tényleges egészséges	90.5%	9.5%
Tényleges depressziós	10.9 %	89.1%

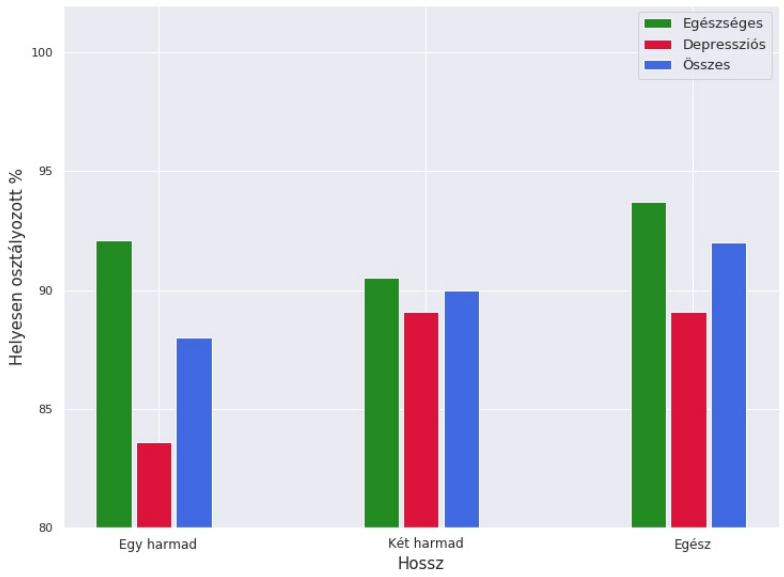
3. táblázat. A két harmad felvételen kapott tévesztési mátrix

Az egész felvételen végzett kísérletnél a C érték 10-nek lett választva az algoritmus által. A kiválasztott jellemzők ugyanazokból a kategóriákból kerültek ki, mint a két harmad felvételen végzett jellemzőválasztás során (MFCC, chromagram, mel-skálás spektrogram, number of voice breaks, tonal centroid). Az eredmények a 4-es számú táblázatban láthatóak.

	Osztályozott egészséges	Osztályozott depressziós
Tényleges egészséges	93.7%	6.3%
Tényleges depressziós	10.9 %	89.1%

4. táblázat. Az egész felvételen kapott tévesztési mátrix

A különböző tévesztési mátrixokból kivehető, hogy a felismerés pontossága alapvetően javul, ha hosszabb felvételt használunk (ami várható is volt). A két harmad felvételt használva növelni lehetett a depressziósok helyes osztályozását az egy harmad felvételhez képest. Bár megnőtt a hibásan depressziósnak osztályozottak száma, ilyen estekben fontosabb, hogy a ténylegesen betegeket minél jobban felismerjük (továbbá egészében a két harmad felvétel 2%-kal pontosabb volt az egy harmadnál, mint ahogy azt láthattuk az első táblázatban).



2. ábra: Az osztályozás pontossága felvétel-hossz és osztály szerint

Az egészségesek helyes felismerése az egész felvétel használatával javult fel. Az egész felvételen a depressziósok felismerése maradt a két harmad felvétel szintjén, de az egészségeseknek a felismerési pontossága az egy harmadhoz képest is nőtt. Ezekből az adatokból érdekes felvetések is felállíthatóak – bár sok tényező játszik közre, az eredmények alapján feltűnik, hogy a felvételek első és utolsó harmada (eleje és vége) bizonyos okokból kifolyólag az egészségesek felismeréséhez volt fontos, a közepe pedig a depressziósokról rejtett több információt.

Mivel a felvételek egész hossza mindössze 40 másodperc körül mozog és ezzel is 90% körüli pontossággal lehetett következtetni az alanyok állapotára, az egész felvételen elért eredmény az algoritmus és a páciensek igényeit is jó mértékben ötvözi.

5. Összegzés és konklúzió

A cikkben bemutattunk egy, a depresszió osztályozására készített hang-alapú automatikus felismerő rendszert, amely ötvözi az akusztikai jellemzők kinyerését, azoknak a kiválasztását (Fast Forward Selection módszerrel) és a hiperparaméter optimalizációt (Grid Search módszerrel). Az osztályozáshoz szupport vektor klasszifikációt használtunk, radiális kernellel és különböző hiperparaméter-

kombinációkkal. Mindezek az eljárások a kisebb adatbázisokon használatos k-Fold Cross Validation módszerrel lettek becsülve pontosságra.

A kísérletek során azt vizsgáltuk, hogy a felvételek hossza hogyan befolyásolja a rendszerünk teljesítményét. Az adatbázisunkban található eredeti felvételek három részre lettek osztva, mondatok félbeszakítása nélkül, fonéma szegmentálás segítségével. Ebből lettek kialakítva az egy harmad, a két harmad és az egész felvétel csoportjai. A teljesítmények becslésére tévesztési mátrixokat használtunk, amelyek kimutatták a helyesen és hibásan becsült minták százalékát az egészséges és depressziós osztályoknál külön is.

A beválasztott jellemzők alapján a legjobban a mel-skálás spektrogram, a mel-frekvenciás kepsztrális együtthatók, a chromagram, a tonal centroid, valamint a number of voice breaks adja meg a helyes osztályozáshoz szükséges információkat.

A teszt eredmények azt mutatták, hogy minél hosszabb felvételt használtunk, a pontosság teljességében nőtt, két-két százalékkal. A legjobb eredményt az egész felvételen értük el, ahol is 92% pontossággal tudtuk az egészségi állapotot megbecsülni. A két osztály klasszifikációs eredményeit külön-külön tekintve érdekes fejleményeket figyelhettünk meg, miszerint a felvételek eleje és vége leginkább az egészségesek helyes felismeréséhez járult hozzá, a közepe pedig a depressziósokról rejtett több információt. Ennek a felvetésnek a helyességét és esetleges hatását következő munkákban érdemes lehetne komolyabban megvizsgálni.

Köszönetnyilvánítás

A K128568 számú projekt a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, a K pályázati program finanszírozásában valósult meg.

Irodalomjegyzék

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: Detecting depression: a comparison between spontaneous and read speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7547–7551. IEEE (2013)

Andrade, L., Caraveo-Anduaga, J.J., Berglund, P., Bijl, R.V., Graaf, R.D., Vollebergh, W., Dragomirecka, E., Kohn, R., Keller, M., Kessler, R.C., et al.: The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. International journal of methods in psychiatric research 12(1), 3–21 (2003)

Association, A.P., et al.: Diagnostic and statistical manual of mental disorders. BMC Med 17, 133–137 (2013)

Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F.: Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. Journal of personality assessment 67(3), 588–597 (1996)

Boersma, P., et al.: Praat, a system for doing phonetics by computer. Glot international 5 (2002)

Buschmeier, H., Wlodarczak, M.: TextGridTools: A TextGrid processing and analysis toolkit for Python. In: Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013) (2013)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Communication 71, 10–49 (2015)

Hays, R.D., Wells, K.B., Sherbourne, C.D., Rogers, W., Spritzer, K.: Functioning and well-being outcomes of patients with depression compared with chronic general medical illnesses. Archives of general psychiatry 52(1), 11–19 (1995)

Jadoul, Y., Thompson, B., De Boer, B.: Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics 71, 1–15 (2018)

Kiss, G., Sztahó, D., Vicsi, K.: Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features. In: 2013 IEEE 4th international conference on cognitive infocommunications (CogInfoCom). pp. 579–582. IEEE (2013)

Kiss, G., Vicsi, K.: Physiological and cognitive status monitoring on the base of acoustic-phonetic speech parameters. In: International conference on statistical language and speech processing. pp. 120–131. Springer (2014)

Kiss, G., Vicsi, K.: Comparison of read and spontaneous speech in case of automatic detection of depression. In: 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000213–000218. IEEE (2017)

Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14, pp. 1137–1145. Montreal, Canada (1995)

Lépine, J.P., Briley, M.: The increasing burden of depression. Neuropsychiatric disease and treatment 7(Suppl 1), 3 (2011)

Mao, K.: Fast orthogonal forward selection algorithm for feature subset selection. IEEE Transactions on Neural Networks 13(5), 1218–1224 (2002)

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference. vol. 8 (2015)

Nilsonne, A.: Speech characteristics as indicators of depressive illness. Acta Psychiatrica Scandinavica 77(3), 253–263 (1988)

Python, J.: Python programming language. In: USENIX Annual Technical Conference (2007)

Rutowski, T., Harati, A., Lu, Y., Shriberg, E.: Optimizing speech-input length for speaker-independent depression classification. Proc. Interspeech 2019 pp. 3023–3027 (2019)

Strakowski, S., Nelson, E.: Major Depressive Disorder. Oxford University Press (2015)

Williams, J.B.: A structured interview guide for the hamilton depression rating scale. Archives of general psychiatry 45(8), 742–747 (1988)

Poszter, laptopos bemutató

FORvoice 120+: magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra

Beke András, Szaszák György, Sztahó Dávid

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
1117 Budapest, Magyar tudósok körútja 2.
beke.andras@gmail.com, {szaszak, sztaho}@tmit.bme.hu

Kivonat: A jelen tanulmányban elsőként kerül bemutatása a FORvoice 120+ magyar nyelvű kriminalisztikai célú utánkövetéses adatbázis. A FORvoice célkitűzése egy kriminalisztikai szempontból megbízható, követéses, reprezentatív beszélői adatbázis elkészítése magyar nyelven. Az adatbázis vizsgálati anyagot biztosít a magyar nyelven történő kriminalisztikai fonetikai kutatásokhoz, illetve a törvényszéki hang-összehasonlító rendszerek fejlesztéséhez és kiértékeléséhez. Az adatbázis 120 beszélő (60 női és 60 férfi) felvételét fogja tartalmazni. A felvételek szigorú protokoll szerint történnek, amelyek követik a nemzetközi irányokat. A FORvoice lehetőséget biztosít, hogy azon akusztikai, fonetikai, nyelvészeti, beszédtechnológiai kutatásokat végezhessenek, külön tekintettel az adatközlő egyéni beszéd tulajdonságára, továbbá a törvényszéki hang-összehasonlító rendszerek fejlesztéséhez és kiértékeléséhez, új, egyéni akusztikai-fonetikai jellemzők megállapításához.

1 Bevezetés

Az elmúlt évtizedekben egyre növekedett az a szakmai erőfeszítés, amely azt a lehetőséget vizsgálja, hogy vajon a beszéd milyen egyéni, személyhez köthető tulajdonságai hordozzák a beszélőspecifikus jellemzőket (vö: Tomi Kinnunen 2018: An Overview of Text-Independent Speaker Recognition: from Features to Supervectors). Mindezen kérdések még nagyobb figyelmet kapnak a törvényszéki munka során, ahol a cél egy kérdéses mintán hallható személy kilétének megbízható azonosítása objektív, statisztikai, megismételhető módszerek segítségével, ahogy azok a DNS-tesztek módszertanában ismeretesek. A beszélők személyének felismeréséhez szükséges egy olyan magyar nyelvű, sok beszélőt tartalmazó adatbázis, amely kriminalisztikai céloknak megfelel, és amely lehetőséget biztosít kriminalisztikai fonetikai, nyelvészeti és beszédtechnológiai kutatások elvégzéséhez. A jelen tanulmány egy ilyen adatbázis fejlesztését mutatja be, amely legalább 120 beszélőt tartalmaz, szigorú protokoll mentén rögzített, az egyes beszélőktől időben eltérő hangmintákat tartalmaz, illetve különböző beszédtípusokat. Az adatbázis jelentősége igen nagy, mivel lehetőséget biztosít a beszélők beszédének személyspecifikus jellemzőinek vizsgálatára. Ugyanakkor az adatbázis és az azon elvégzett kutatássorozat nem csak a kutatók számára hasznos, hanem a társadalom szá-

mára is, mivel egy ilyen adatbázis lehetőséget biztosít a rendőri szerveknek, nemzetbiztonsági szerveknek, hogy a törvényszéki munka során használt rendszereket megbízhatóbbá tegyék, újakat fejlesszenek.

A kriminalisztikai hang-összehasonlítás során arra keressük a választ, hogy az időben korábban a hivatalos szervekhez érkezett hangminta (jellemzően telefonos spontán beszédet tartalmazó minta) ugyanattól a személytől származik-e, akitől később az eljárás során interjú szituációban vettek hangmintát (jellemzően nem spontán stúdió hangminőségű hangminta). A kriminalisztikai hang-összehasonlításkor a beszédmintákat akusztikailag elemzik, és ezen az elemzésen alapulva mutatják be, hogy a hasonlóságot milyen mértékben növeli vagy csökkenti a keletkezett bizonyíték ezzel segítve a bírói döntési mechanizmust.

A kriminalisztikai tudományban bekövetkezett paradigmaváltás (Saks és Koehler, 2005) előtt az elemzés során elégséges volt csak a két hangminta akusztikai összevetése, vagyis annak prezentálása, hogy az adott akusztikai jegy (pl. alaphangmagasság) nagy hasonlóságot vagy különbséget mutat-e. Ugyanakkor belátható, hogy fennállhat az az eset is, hogy egy populációból véletlenszerűen vett két beszélőnél ugyanezt a hasonlóságot vagy különbséget találnánk. A kérdés tehát az, hogy az adott akusztikai jegy(ek) mennyire hasonló(ak), illetve mennyire tipikus(ak) az adott egyénre, illetve a populációra nézve. Ezt a kérdés oldotta fel a kriminalisztikai tudományban bekövetkezett paradigmaváltás (Saks és Koehler, 2005), amely a bizonyíték kiértékelésében, illetve prezentálásában hozott változásokat, és amely forradalmasította a kérdéses és a gyanúsítottól származó minta összehasonlításának módszertanát (vö. DNS-profil stb.). Az új paradigmát a valószínűségi-arány keretrendszer (likelihood-ratio framework) kvantitatív implementációjaként lehet jellemezni, amely az eredmények megbízhatóságának kvantitatív úton történő kiértékelését biztosítja. A likelihood-ratio framework során két alapvető hipotézis kell megvizsgálni. A jogalkalmazó által az igazságügyi szakértőnek feltett alapkérdés: „Mekkora valószínűséggel származik a kérdéses minta a gyanúsított személytől?”, illetve az ún. ellenhipotézis: „Mekkora valószínűséggel származik a kérdéses minta az adott népességből véletlenszerűen kiválasztott másik személytől?”. Mindkét, ún. posterior valószínűség kiszámításához a Bayes-elv alapján hipotézisenként két valószínűségi értéket kell kiszámolni, majd a kapott valószínűségeket egymással elosztani. Az igazságügyi kérdés a likelihood framework tükrében tehát az, hogy „Mennyiszer tűnik valószerűbbnek, hogy a megfigyelt különbségek az ismert és a kérdéses minták között azt a feltételezést támogatja, hogy a kérdéses mintának és az ismert mintának azonos az eredete, mint azt a feltételezést, hogy az eredete különböző?” (lásd bővebben Geoffrey Stewart Morrison munkáit).

Ahhoz, hogy a LR keretein belüli kísérleteket elvégezhessük, szükséges egy olyan adatbázis, amely az új paradigma alapfeltevéseinek megfelel (Morrison és mtsai, 2012): 1):

(i) minden beszélőtől időben eltérő mintákat kell rögzíteni (hasonlóság modellezése),
(ii) sok beszélőt kell tartalmaznia lehetőleg a populációra reprezentatíven (tipikusság modellezéséhez),
(iii) különböző módon rögzített hangmintákat kell felvenni (un. channel mismatch kompenzálására, pl. telefonos vagy stúdió minőségű),

(iv) egy beszélőtől különböző beszédtípusokat kell rögzíteni a beszédstílus különbségeiből fakadó beszélőn belül is megjelenő eltérések kompenzálására (speech style mismatch compensation).

Mindezen kihívásoknak megfelelően terveztük meg a jelen tanulmányban bemutatott FORvoice adatbázist, amely a jelenleg elérhető hazai adatbázisok között egyedülálló (vö. MTBA (Vicsi és mtsai, 2004), MRBA (Vicsi és mtsai, 2004), BABEL (Vicsi és Vig, 1998), BEA (Gósy és mtsai, 2012)).

2 Célkitűtések

A fejlesztett FORvoice adatbázis a következők célkitűzések mentén épül fel. Célunk egy olyan adatbázis létrehozása, amely

(i) illeszkedik a kriminalisztikában bekövetkezett új paradigmaváltásban megfogalmazott kritériumokhoz, így a rajta végzett elemzések fontos alapkövei lehetnek a törvényszéki hang-összehasonlító rendszerek fejlesztéséhez és kiértékeléséhez, új, egyéni akusztikai-fonetikai jellemzők megállapításához,

(ii) annotált és lekérdezhető, így lehetőséget biztosít a szakemberek számára, hogy azon akusztikai, fonetikai, nyelvészeti, beszédtechnológiai kutatásokat végezhessenek, külön tekintettel az adatközlő egyéni beszéd tulajdonságaira, továbbá

(iii) olyan alap adatbázis legyen, amelyen új kutatási irányokat lehessen megvalósítani a fonetikában, a beszédtechnológiában és olyan kutatási kérdések megválaszolására adjon alapot, amelyeket korábban nem lehetett tanulmányozni magyar nyelven (pl. a beszélőn belüli és a beszélők közötti variancia szisztematikus elemzése hosszabb időtávon, stb.).

3 Anyag, módszer és kísérleti személyek

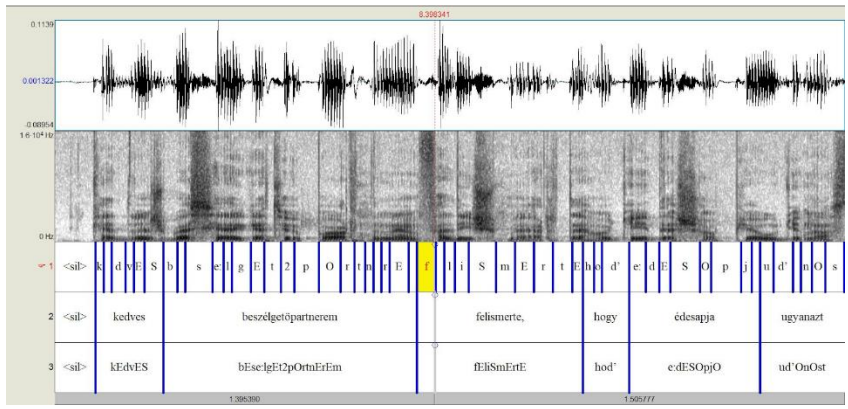
Az adatbázis készítése során a fő szempontok igazodnak a nemzetközi irodalomhoz (Morrison és mtsai, 2012): 1) minden beszélőtől legalább két, időben relatíve távoli felvételt kell tartalmaznia; 2) az egyes személyektől különböző beszédtípusokat kell rögzíteni: alkalmi beszélgetés, irányított beszélgetés, ál-rendőrségi-kihallgatás (monológ formájában); 3) az adatbázisnak meg kell felelnie a kutatások és a kriminalisztikai esetek követelményeinek (a felvételi és adatátviteli csatorna közötti különbségek modellezése). A felvételi és adatátviteli csatorna eltérésének kritériuma utólag kerül modellezésre digitális jelfeldolgozás segítségével.

A FORvoice 120 beszélő hangmintáit fogja tartalmazni (60 női és 60 férfi). A beszélők 18-65 év közötti magyar anyanyelvű személyek, változatos születési hellyel. Minden személytől kétszer készül felvétel, a két alkalom között legalább két hét különbséggel. A felvételek előtt az alanyok írásban nyilatkoznak a hangjuk rögzítéséhez való hozzájárulásukról. A hangfelvételek mellett rögzítjük a beszélők fittségi és stressz szintjét (önbevallásos alapon), illetve, hogy dohányoznak-e, valamint azt, hogy van-e hangképzést befolyásoló betegségük.

A felvételek párosával történnek egy csendes irodai szobában. A beszélő párok 2-3 m távolságra ülnek egymástól (1. ábra). A hangrögzítés USB-s külső audió interfész és fejmikrofonok segítségével történik, a kezdeti jó minőség biztosítása érdekében (2. ábra). A felvétel során a beszélőknek négy feladatot kell elvégezniük:

(i) *Szabad párbeszéd*. Teljesen szabad párbeszéd, kötöttségek nélkül.
(ii) *Célzott információcsere*. A beszélőpárosok egy-egy hibás terméklista számukra 4-4 olvashatatlan soraiban lévő információit kell megszereznie a beszélgetőtárstól.
(iii) *Monológ*. A beszélőknek az előző napjukat kell elmesélniük tárgyilagosan.
(iv) *Állandó válaszú kérdések*. A beszélőknek 5-5 rögzített kérdésre kell 'Nem emlékszem', illetve 'Nem kívánok válaszolni' választ adniuk.

1. ábra. A hangfelvétel helyszíne



2. ábra. Spektrogram és szegmentálás minta egy elkészült jó minőségű felvételről

A FORvoice jelenleg 60 beszélő teljes protokolját tartalmazza. Elkészült a hanganyag szó- és hangszintű átiratozása is. Mindemellett további kiegészítésként az intonációs frázisok jelölése is megvalósult.

4 A FORvoice-on tervezett kutatások

A korai kutatások során elsőként az alaprendszereket (baseline) szeretnénk rögzíteni és publikálni, vagyis a jelenleg széles körben használt i-vector alapú beszélőfelismerő algoritmusok kiértékelése történik meg a FORvoice-on.

Emellett leíró jellegű temporális akusztikai-fonetikai elemzéseket is tervezünk az egyéni beszédjellemzők vizsgálatára. Későbbiekben vizsgáljuk a kriminalisztikai szempontú beszélőspecifikus akusztikai-fonetikai paramétereket. Elemezzük, hogy az új beszélőjellemzők az általunk létrehozott rendszer eredményeit milyen módon javítják. Alapvetően három nagyobb területen tervezünk kísérleteket végezni: i) temporális jellemzők, ii) prozódiai jellemzők, iii) mély neurális hálózatokon alapuló jellemzők. A kutatás során olyan akusztikai paramétereket vizsgálunk, amelyek dinamikus változása jól tükrözi az artikulációs szervek egyéni működését, ilyen módon beszélőspecifikus jellemzőként működnek, valamint jól vizsgálhatók az agglutináló típusú nyelv esetében. A temporális vizsgálatok során az egyes hangfelvételekből különböző típusú időzítési információk kinyerését végezzük el, amelyek alapján elemezzük a beszélők közötti és az egyes beszélőkre jellemző variancia mértékét. Kísérletet tervezünk a mély neurális hálók alkalmazására a törvényszéki hang-összehasonlító rendszer jellemzőkinyerésére. A mély neurális hálók segítségével lehetőség nyílik további lokális jellemzőkinyerésre is. Egy ilyen technika a mély hálók utolsó rejtett réteg kimeneteinek használata (Bacchiani és Rybach, 2014). Több tesztben magukat a hálók kimeneteit is sikerrel alkalmazták (pl. Senior és mtsai, 2014); ugyanakkor ahhoz, hogy a már bejáratott, gaussi modellezési technikákat minél nagyobb pontossággal lehessen alkalmazni az így kinyert jellemzőkön, szükséges lehet azok transzformálása (Zhang és Woodland, 2014).

Az egyénre jellemző intonációs vagy hangsúlyozási mintázatok automatikus analízise mellett a szövegtagolás és központozás egyéni specifikumait is vizsgájuk: menynyiben tárhatók fel egyénre jellemző mintázatok, milyen konfidenciával. Automatikus eszközök használatára törekszünk (prozódiai esemény detektálókra és intonációs osztályozókra), amelyeket a kísérleti rendszerbe is integrálunk. Vizsgáljuk továbbá, hogy milyen módon lehet az egyes akusztikai-fonetikai jellemzőket egymással kombinálni, és ezeknek milyen hatás van a rendszer kimenetére. Elemezzük tovább, hogy a különböző akusztikai jellemzőkkel kapott kimeneti értékeket (valószínűségi-arány érték: Likelihood Ratio Score) milyen módon lehet kombinálni a rendszer végleges eredményének javításához.

Az i-vektorok számításának egy neurális hálózattal megvalósított enkóder-dekódereken alapuló alternatív eljárását is kidolgozzuk, amitől a beszélők mélyebb jellemzését, leírását várjuk. Az enkóder bemenetére a beszélőtől származó, a dekóder kimenetére univerzális vagy beszélőfüggetlen mintát téve a rejtett rétegen kinyerhető a tömör,

átlaghanghoz viszonyított beszélőreprezentáció. Tervezzük, hogy ezzel a módszerrel végzünk kísérleteket az adatbázison mérve annak performanciáját.

5 A FORvoice elérhetősége

Az adatbázist előreláthatólag 2022-ben fogjuk nyilvánosan elérhetővé tenni.

6 Összegzés

A FORvoice fontos mérföldköve lehet a magyarországi kriminalisztikai tudományának. Lehetőséget biztosít a beszélőre specifikus akusztikai-fonetikai, nyelvészeti jegyek vizsgálatához magyar nyelven a fonetikai, a nyelvészeti és a beszédtechnológiai szakemberek számára. Társadalmi hasznossága kiemelkedő, mivel ez lesz az első olyan magyar beszédkorpusz, amely kriminalisztikai szempontoknak megfelel és így standardizálttá válik.

Az adatbázis lehetőséget biztosít mind a tudományos szakmai közönség, mind pedig a rendészeti szervek számára, a törvényszéki hang-összehasonlítás módszertanának vizsgálatára, illetve általánosságban beszélőfelismerő rendszerek kifejlesztésére és kiértékelésére. A FORvoice adatbázison végzett kutatások új ismereteket nyújtanak az adatközlő egyéni beszéd-sajátosságairól, és alapot adnak további nyelvészeti, beszédtechnológiai vizsgálatokhoz. A fejlesztendő adatbázis – szigorú protokolljának, felvételi módszertanának, az annotálásnak és mennyiségi jellemzőinek köszönhetően (sok beszélő, utánkövetéses eljárás, beszélőnként több felvétel, különböző beszédtípusok) – kiválóan alkalmazható elsősorban

- a bűnügyi beszélőazonosításban,
- az automatikus beszédfelismerő rendszerekben,
- a beszédszintézisben és beszédfelismerésben a beszélőadaptációban,
- de tágabb felhasználási területként minden követéses adatot igénylő kutatásban vagy fejlesztésben – így a fonetika, a beszéd alapú egészségügyi diagnosztika, stb.

A tervezett adatbázis nemzetközi tudományos értéke mellett, jelentős nemzeti értéket is képvisel.

Köszönetnyilvánítás

Az FK128615 számú projekt a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, az FK pályázati program finanszírozásában valósult meg.

Hivatkozások

Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 230-234). IEEE. (2014)

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T. E., Beke, A., Neuberger, T., Nikléczy, P.: BEA: Beszélt nyelvi adatbázis [BEA – A Hungarian Spontaneous Speech Database]. In Gósy, M. (ed.): Beszéd, adatbázis, kutatások. Budapest: Akadémiai Kiadó. pp. 9-24. (2012)

Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40. (2010)

Morrison, G. S., Rose, P., Zhang, C.: Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Australian Journal of Forensic Sciences, 44(2), 155-167. (2012)

Saks, M. Koehler, J.: The coming paradigm shift in forensic identification science. Science Magazine, 309, 892-895 (2005)

Vicsi, K., Kocsor, A., Teleki, Cs., Tóth, L.: Beszédadatbázis irodai számítógép-felhasználói környezetben, Second Conference on Hungarian Computational Linguistics (MSZNY 2004), Szeged, 2004. po. 315 (2004)

Vicsi, K., Vig, A.: First Hungarian Speech Database. Beszédkutatás ’98. pp. 163–177. (1998)

Zhang, C., Woodland, P. C.: Standalone training of context-dependent deep neural network acoustic models. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5597-5601). IEEE. (2014)

Longitudinális korpusz magyar felnőtt adatközlőkről

Gráczi Tekla Etelka[1,2], Huszár Anna[1], Krepsz Valéria[1],
Száraz Bettina[1], Damásdi Nóra[3], Markó Alexandra[2,4]

[1] Nyelvtudományi Intézet, Pf. 360,
H-1394, Budapest, Magyarország,
{graczi.tekla.etelka, huszar.anna, krepsz.valeria,
szaraz.bettina}@nytud.hu
[2] MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport, Múzeum krt. 4/A,
H-1088, Budapest, Magyarország
marko.alexandra@btk.elte.hu
[3] ELTE Bárczi Gusztáv Gyógypedagógiai Kar, Ecseri út 3.,
H-1097, Budapest, Magyarország
damasdi.nora@barczi.elte.hu
[4] ELTE BTK Alkalmazott Nyelvészeti és Fonetikai Tanszék Múzeum krt. 4/A,
H-1088, Budapest, Magyarország

Kivonat: Az ún. longitudinális korpusz rögzítőinek célja felnőtt beszélők követéses vizsgálata a beszéd különféle sajátosságainak a tekintetében. Az adatközlők magyar anyanyelvűek, akiknek a hanganyagát először a BEA adatbázisban rögzítették, majd 10-11 évvel később a longitudinális korpusz módszertanával is felvételeket készítettek velük. A korpusz beszéd kutatók számára hozzáférhető lesz. A tanulmány ismerteti a korábbi longitudinális kutatásokat, amelyek a jelen korpusz alapjául szolgáltak a módszertan kialakítása szempontjából, valamint bemutatja a folyamatban lévő korpuszépíési munkálatokat.

1 Bevezetés

A beszéd variabilitásában az egyik központi tényező a beszélő életkora. Ez számos tanulmányban szerepel faktorként, de leggyakrabban keresztmetszeti vizsgálatokban, mivel nagyobb számú adatközlő több évnyi távlatban történő utánkövetése nehézkes, gyakran megvalósíthatatlan feladat. A jelen tanulmányban egy olyan longitudinális korpusz tervét, munkálatait mutatjuk be, amely a Nyelvtudományi Intézet Fonetikai Osztályán készül. A projekt egy NKFI-pályázat (FK-128814) keretében zajlik, amelynek több központi kérdése mellett a beszéd mintegy egy évtized alatt történő, beszélőn belüli változásának feltárása a célja. A longitudinális korpusz és elemzések tehát a beszéd 10-11 év alatti változását ragadják meg. Ez az időtartam olyan közepes intervallum, amely elegendő lehet, hogy mind az elsődleges (kronológiai), mind a másodlagos öregedés (pl. beszédszokások, vokális terhelés) hatásai megmutatkozzanak. A készülő felvételeket és annotációkat a beszédkutatók számára hozzáférhetővé fogjuk tenni.

A jelen tanulmányban a korpusz kiépítésének elméleti hátterét, a folyamatban lévő munkálatokat, illetve az eddigi felvett anyagot mutatjuk be.

1.1 Életkoralapú kutatások

A különböző tudományos elemzések és kutatások esetében központi kérdés a megfelelő módszertani eljárás megválasztása. Azokban az esetekben, ahol az életkor vagy az idő múlása szerepet játszik, kétféle elemzési eljárás alkalmazható: keresztmetszeti vagy követéses, azaz longitudinális vizsgálat. A beszédhez kapcsolódó elemzésekben ezek a fogalmak elsőként a szociolingvisztikában jelentek meg, ahol látszólagosidő-vizsgálatként, illetve valóságosidő-vizsgálatként szerepeltek (Labov, 1994, 2001; Trudgill, 1974). Az első esetben egy közösség fiatal és idős tagjainak vizsgálatával, a második esetben ugyanazon beszélő időbeni követésével, azaz a történés valós folyamatában elemzik a csoportok közötti különbségeket, és ez alapján megkísérlik azonosítani az idő múlásának a különböző tényezőkre gyakorolt hatását.

Az anyanyelv-elsajátítási kutatásokban és a különböző patológiás esetek elemzésében már korábban is elterjedt volt a longitudinális módszertan alkalmazása. Ide tartozik például a demencia (pl. Lee és mtsai, 2011), az Alzheimer-kór (pl. Blair, 2007), a Parkinson-kór (Ash és mtsai, 2017) beszédre gyakorolt hatásának követéses vizsgálata vagy a nyelvi és beszédzavarok gyógyítására használt terápiák hatékonyságának vizsgálata (pl. Powell és mtsai, 1989, Misono, 2016) is.

A betegségek leírása mellett az életkori jellemzők is gyakran kerülnek a vizsgálatok középpontjába. Eddig azonban csupán néhány követéses elemzés látott napvilágot (pl. Reubold és mtsai, 2010; Hunter és mtsai, 2012). Ennek az egyik oka, hogy jelentős módszertani nehézséget jelent a kulturális, nemi, pszichoszociális beszédjellemzők vagy éppen egyes betegségek beszédbeli tüneteinek és az idősödő hang sajátosságainak elválasztása. Kérdés tehát például az, hogy az adott beszélő beszédében aktuálisan egy adott paraméterben mért csökkenő vagy növekvő tendencia valamely betegséggel magyarázható-e, vagy valóban az életkor előrehaladtával bekövetkező sajátosság.

Az egyes szervek öregedése eltérő tempóban zajlik, így az egyes beszédképzőszervek változása is. A legjelentősebb változások a felnőttek esetében idős korban történnek, amit azonban számos tényező befolyásol (vö. Liu-Su, 2017). Ugyanakkor a változás már korán, 30 éves kor körül megindul: a tüdőerek merevsége már 30-35 éves kor után kimutatható, a tüdőkapacitás kb. 30 éves kortól csökken (összefoglalás: Lalley, 2013). Az öregedés folyamatát két tényezőre szokás osztani, az elsődleges és a másodlagos öregedésre (Busse, 2002). Az elsődleges öregedés az életkor előrehaladtával (kronológiailag) végbemenő, genetikailag meghatározott változásokat, míg a másodlagos öregedés a betegségek, környezeti hatások, káros szokások hatására bekövetkező változásokat foglalja magában. Az egypetéjűiker-vizsgálatok eredményei szerint nagyobb arányban a genetika, azaz az elsődleges öregedés tényezői, és csupán kisebb arányban az életvitel és más tényezők, külső hatások, azaz a másodlagos öregedés felel az idősödés folyamatának minőségéért (Guyuron és mtsai, 2009). Az életkor előrehaladtával bekövetkező változások a beszédjellemzőkben is megjelennek, például az alapfrekvencia mélyülésével, a tüdőkapacitás romlásával, a fokozatosan megváltozó agyműködéssel, ami a beszédtervezést és -kivitelezést is érintheti stb. Az öregedésen túl természetesen további tényezők, pl. a szociális jellemzők is hatással vannak a beszédjellemzőkre.

1.2 A longitudinális és a keresztmetszeti vizsgálatok eltérései

A keresztmetszeti elemzések egy adott időszakban az életút különböző fázisaiban lévő (egyéb szempontok szerint homogén csoportot alkotó) egyének vizsgálatát jelentik, míg a követéses vagy hosszmetszeti, azaz longitudinális vizsgálatok az egyéni változások elemzése az idő/az életkor előrehaladtával vagy azonos adatközlő(i csoport) analízise hosszabb időintervallumon keresztül. A keresztmetszeti elemzések esetében az eredmények az interindividuális, beszélők közötti különbségeket mutatják be, és minden beszélő „egy értékkel" jellemezhető. Ezzel szemben a longitudinális elemzések az intraindividuális, beszélőn belüli különbségeket ragadják meg, így egy beszélő „több értékkel" is jellemezhető. A keresztmetszeti elemzések során nem a változás vizsgálata történik, arra csupán közvetve következtethetünk. Ezzel szemben a hosszmetszeti elemzés során a valódi változás vizsgálata történik, de számos egyéb, kiküszöbölhetetlen hatás is befolyásol(hat)ja az értékek alakulását. Míg az első esetben nagyobb adatközlői csoport vizsgálata történik, addig a longitudinális vizsgálatokban általában egy adatközlő vagy kisebb beszélő csoport vizsgálata lehetséges, mivel a beszélők visszahívása/újbóli felkeresése nehézkes. Végül a hiba típusa is eltérő a két esetben: A keresztmetszeti kutatásokban a lehetséges hiba a szisztematikus hiba vagy a véletlenszerűen jelentkező hiba, azaz egyetlen generációt (ebben az esetben egyetlen adatközlői csoportot) érintő hiba (más néven kohorszeffektus; Jacob-Ganguli, 2016), míg a longitudinális vizsgálatokban kiküszöbölődik a kohorszhatás, de az esetek nagy részében alacsony adatközlőszám miatt az egy-egy beszélőre jellemző tendenciák felerősödhetnek. A longitudinális vizsgálatok esetében az is gyakori, hogy specifikus csoportot vizsgálnak. Ilyen például a rádióbemondók beszéde, akik sok esetben a hétköznapi beszélőkre nem jellemző beszédtréningen esnek át.

1.3 Nemzetközi longitudinális korpuszok és vizsgálati eredmények

A módszertani nehézségekből adódóan korábban részben kevés, felnőttcsoporton végzett longitudinális vizsgálat látott napvilágot, másrészt ezek általában vagy egy-egy személy (gyakran professzionális beszélő) vagy speciális adatközlői csoport adatait dolgozták fel. Emellett nagyban nehezíti a longitudinális elemzések összevetését, hogy az egyes vizsgálatok eltérő módszertani eljárásokat alkalmaztak. Az alábbiakban összefoglaljuk az eddigi jelentős eredményeket.

Az egyik sokat vizsgált korpusz II. Erzsébet angol királynő karácsonyi beszédeinek gyűjteménye (pl. Harrington és mtsai, 2007; Reubold és mtsai, 2010). A felvételek az adott időszakok technológiai sajátosságainak megfelelően jó minőségűek. Emellett alkalmassá teszi az anyagokat a longitudinális összevetésre, hogy minden esetben azonos a beszédhelyzet (karácsonyi köszöntő, évértékelő), azonos periódusonként vették fel a hanganyagokat (minden évben karácsonykor), illetve minden felvétel tartalmaz legalább egy állandó mondatot *(I wish you a peaceful and very happy christmas.).* Felmerül azonban a kérdés ebben az esetben is, hogy egyetlen beszélő beszédprodukciójának elemzése alapján milyen következtetések vonhatók le. Harrington és munkatársai (2000b) II. Erzsébet beszédeiben a szóhangsúlyi helyzetű monoftongusok formánsértékeit vizsgálva megállapították, hogy a fenti időtartomány alatt vertikálisan tágult, horizontálisan valamelyest szűkült a királynő magánhangzótere. Ugyanakkor ők maguk

ebben és másik tanulmányukban (2000a) is leírják, hogy ennek az életkori változáson túl más szociofonetikai oka is lehet, mégpedig az, hogy az uralkodó kiejtése feltehetően a fiatalabb és szegényebb rétegéhez közeledett valamelyest.

Harrington és munkatársai (2007) egy további kutatásuk során négy adatközlő beszédében elemezték az akusztikai szerkezet alakulását az idő előrehaladtával: II. Erzsébet, Margaret Lockwood színésznő és egy ausztrál, valamint egy új-zélandi férfi beszédét vették górcső alá. Az alapfrekvencia, valamint a nem hangsúlyos szótagi svá-realizációk első három formánsának átlagos értékét. II. Erzsébet angol királynő esetében az 1950-es évektől a 2000-es évekig rögzített karácsonyi beszédeit elemezték. Ezek a felvételek átlagosan 5 perc körüliek (1–8 perc). A felvételek 26 éves korától egészen 76 éves koráig fedik le a királynő beszédét. Mind az alapfrekvencia, mind az első és második formáns esetében csökkenést figyeltek meg az életkor előrehaladtával. II. Erzsébet királynő átlagos alapfrekvenciája a hozzávetőlegesen 50 év mintegy 23%-kal lett alacsonyabb: 267 Hz-ről 208 Hz-re csökkent. A csökkenés mértéke az 1950-es (34+ éves) és az 1990-es években (64+ éves) volt a legnagyobb. Az első formáns értéke 530 Hz-ről 414 Hz-re változott. Ennek a változásnak a mértéke az 1960-as években volt a legnagyobb. A második formáns értéke az 50 év alatt 1796 Hz-ről 1716 Hz-re csökkent. A változás itt kisebb mértékű volt, mint az előző két esetben. Az 1960-as és 1970-es években kisebb mértékű volt a csökkenés, mint a többi évtizedekben. A harmadik formáns esetében kismértékű növekedés volt megfigyelhető az első és az utolsó év között, de nem jelentős, illetve nem lineáris a változás. Margaret Lockwood angol színésznő (született: 1916) beszédét egy 1951-es és egy 1980-as interjúja alapján vetették össze. Az első felvétel 5 és fél perces, a második 12 perces. A királynő értékeihez hasonlóan az f0 és az F1 csökkent, a F2 és a F3 azonban valamelyest emelkedett. Ezen túl egy ausztrál (40 és 79 éves korában) és egy új-zélandi férfi beszélő (36 és 73 éves korában) felvételeiben eltérő tendenciát találtak. Az ausztrál férfi esetében az f0, F1, F2 a királynő értékeihez hasonlóan csökkent, a F3 alig emelkedett. Az új-zélandi beszélő esetében mind a négy akusztikai paraméter csökkenést mutatott az életkor előrehaladtával. Mind a négy beszélő esetében az alapfrevencia mutatta a legnagyobb változást, és mind a négy esetben csökkent. A formánsértékek ugyan változást mutattak, de az F1-f0 és az F3-F2 Barkban számított különbsége alapján azt feltételezik a szerzők, hogy a percepció számára megőrződik a magánhangzó minősége. Felteszik a kérdést, hogy a formánsok változása mennyiben függ össze az életkori fiziológiás változásokkal és mennyiben az f0 változásának kompenzációja, hogy a magánhangzó-minőség fenntartható legyen az észlelet számára.

Egy további longitudinális vizsgálatban Reubold és mtsai (2010) ismét II. Erzsébet királynő karácsonyi beszédeit és Margaret Lockwood angol színésznő rádióinterjúit elemezték, de további anyagokat, Margaret Thatcher beszédeit, illetve Roy Plomley és Alistair Cooke angol rádióbemondók anyagait is bevontak. Ismét a hangsúlytalan szótagi svákban mérték az alapfrekvencia és az F1 változását 29-35 év távlatában. Mindkét jellemzőben csökkenést találtak, ugyanakkor Cooke 80-as éveitől mindkét érték emelkedést mutatott. Elemezték azt is, hogy az életkor észlelete függ-e az f0 és az F1 változásaitól. Az eredmények a korábbi, Harrington és mtsai 2007-es munkájának eredményeit megerősítették, és a két tanulmány eredményeit összegezve arra hívják fel a figyelmet, hogy az idős és fiatal beszélők beszédében a formánsértékek összevetésekor nem csak a diakrón változást, hanem az életkori fiziológiás jellemzőket is figyelembe kell venni.

Vannak kutatások, melyek meglévő felvételek szereplőinek anyagát rögzítik újra, így kisebb longitudinális korpuszokat létrehozva. Russell és munkatársai (1995) az Ausztrál Nemzeti Film- és Hangarchívum 1945-ös felvételeinek 15 adatközlőjét (18 és 25 év közötti nők) keresték fel, és újra rögzítették a beszédüket 1993-ban. Decoster és Debruyne (2000) 20, 23 és 42 év közötti belga rádióbemendó eredeti és 30 év után megismételt felolvasását vetette össze az alapfrekvencia és a zöngekezdési idő jellemzőinek elemzése céljából. Az adatok nem mutattak egységes változást az átlagos alapfrekvenciában, de jellemzően csökkent az érték mind az átlagban, mind a szórásban, míg a zöngekezdési idő emelkedett.

Russel és kollégái (1995), valamint Verdonck-de Leeuw és Mahieu (2004) hétköznapi beszélők utánkövetéses beszédvizsgálatát végezte el. Russelék egyetemi hallgatókat, Leeuwék pedig egy orvosi kutatásban szereplő egészséges kontrollszemélyeket hívtak vissza. Mindkét esetben csak a tervezett beszélők egy részét sikerült újra rögzíteni. Russel és munkatársai eredményeiben 48 év távlatában alapfrekvencia-csökkenés mutatkozott. Verdonck-de Leeuw és Mahieu (2004) öt év után is változást mutatott ki a beszédben. Adatközlőik 50 év feletti férfiak voltak, és a pszichés hatás és a dohányzás okozta változások is jelentősek voltak az életkori hatás mellett.

Az életkor mentén történő vizsgálatokban – mint láttuk – fontos tényező maga az életkor, a bekövetkező fiziológiai változásokkal, a másodlagos öregedés (pl. a dohányzás hatása), vagy épp a nyelv diakrón változása. Quené (2013) az artikulációs tempó elemzését végezte el Beatrix királynő beszédeiben 42 és 74 éves kora között. Az értékek több szempontból is nem várt eredményeket hoztak. Egyrészt a folyamatosan csökkenő átlagos artikulációs tempó az utolsó évtized során ismét emelkedett. Másrészt az utolsó évtizedig egy felolvasás során kis mértékű változással egy emelkedő, majd ereszkedő ívvel volt leírható a tempó variabilitása, míg az utolsó évtizedben a szöveg elejétől a végéig folyamatosan emelkedett. Quené tehát levonja a következtetést, hogy az életkori elvárt lassulással szemben több faktort is figyelembe kell venni. Lehetséges okként felhozza általában a tempó gyorsulását, amely a királynő beszédében is megjelenhet, a beszélő gyakorlottságát – miszerint egyre gyorsabban tudja olvasni a beszédet, és esetleg követni a tempó feltételezett diakrón változását és a szöveg kívánalmait –, illetve megemlíti, hogy a beszéd hossza milyen lehet. Ebben a vizsgálatban is felmerül a kérdés, hogy az uralkodó alkalmazkodik a kisebb presztizsű, fiatal közönség ejtéséhez.

Magyarországon felnőtteken eddig klasszikus értelemben vett longitudinális vizsgálatot nem végeztek, de hosszabb távon több felvételen keresztül elemeztek beszélőn belüli variabilitást.

Gósy (2002) négy női beszélő ejtésében ugyanazon mondat ismétlései alapján 10 hónapon keresztül vizsgálta a magyar magánhangzók időtartamát és formánsértékeit. Eredményei azt mutatták, hogy három beszélő esetében kimutatható a különbség az eltelt idő függvényében. Továbbá az eredmények jellegzetes beszélőn belüli és beszélők közötti eltéréseket támasztottak alá a vizsgált magánhangzók esetében.

Gósy és Krepsz (2015) 7 hónapos periódusban, kéthetenkénti hangrögzítés alapján négy magánhangzó időtartamát, mondatok kiejtési idejét és artikulációs tempóját vizsgálták őt női beszélő ejtésében. Valamennyi magánhangzó a mondat első tartalmas szavának hangsúlyos szótagjában jelent meg (pl.: Kérsz egy falatot az almából?). Az eltelt idő függvényében az artikulációs tempóban és a magánhangzók időtartamában nem volt szignifikáns eltérés kimutatható. Három beszélő az egymást követő felvételeken alig mutattak változást a tempóértékeikben, míg két beszélő artikulációs tempója

nagymértékben variálódott. Beszélőktől függetlenül az [ɔ] és az [eː] időtartamai nagyobb különbségeket mutatnak az egyes felvételek között, mint az [ɛ] és az [o] magánhangzókéi.

Magyar nyelven longitudinális adatbázis még nem készült. Ahhoz, hogy ez megvalósítható legyen, szükség van olyan felvételek rögzítésére, amelyek ugyanazon, nagyszámú beszélőtől eltérő időpontokban, egységes módszertannal, stúdiókörülmények között készülnek.

2 Az épülő Longitudinális korpusz

A jelen tanulmányban bemutatott projekt egyik fő célja egy olyan korpusz kiépítése, mely legalább 40, 10-11 év távlatában utánkövetett beszélő felvételét és annak annotációját tartalmazza. A beszélők esetében célunk, hogy lehetőleg nem kizárólag speciális csoportokból (pl. rádióbemondó) kerüljenek ki.

A tízéves időtartam viszonylag újdonságnak számít a beszéd fonetikai leírásában. A legtöbb longitudinális vizsgálatban 25-30 évnyi vagy hosszabb időtávban rögzített felvételeket elemeztek, mivel feltehetően nagyobb a változások mértéke több idő elteltével, így azok könnyebben mérhetők. Más kutatások rövid távú variabilitást (például 7 hónapot) vizsgáltak, amely során a változásokat nem lehet a felnőttek elsődleges öregedésének tulajdonítani. Ez utóbbi variabilitás akár a másodlagos öregedésből (vokális terhelés, beszédmódok stb.) vagy a beszélőn belüli változatosságból is adódhat. Láthattuk, hogy Verdonck-de Leeuw és Mahieu (2004) öt év elteltével is jelentős változásokat mért. Az ő esetükbena a dohányzás is meghatározó tényező volt. Felmerül azonban a kérdés, hogy 10 év távlatában milyen változások következ(het)nek be a beszédben. A jelen korpusz kiépítésével tehát lehetőség adódhat a beszéd egyes jellemzőinek rendszerszerű elemzésére longitudinális aspektusból.

2.1 Beszélők

A Longitudinális korpusz a BEA-adatbázis (Gósy és mtsai, 2012) anyagára épül. A BEA-adatbázis azon beszélőit keressük meg, akik a jelen projekt adott évéhez képest 10-11 évvel korábban vettek részt beszédfelvételen. A felvételre ismét hajlandó személyek beszédét rögzítjük újra. Az utánkövetés természetesen ebben a keretben sem elérhető a teljes adatközlői létszámra, hiszen többek nem adták meg elérhetőségüket, vagy az megváltozott. Célunk, hogy legalább évi tíz, a négyéves projektidő alatt legalább 40 fő utánkövetéses felvételét elkészítsük.

A tanulmány megírásáig, azaz a projekt első évében 17 beszélő új felvételét tudtuk rögzíteni: 12 férfiét és 5 nőét. Életkoruk az első felvétel idején 19 és 45 év között, az utánkövetéses felvételkor pedig 29 és 55 év között volt.

2.2 Adatlap/anamnézislap

A BEA-felvételek után az adatközlők a nemüket, korukat, iskolai végzettségüket, foglalkozásukat, magasságukat, súlyukat, dohányzási szokásaikat és esetleges beszédhibájukat adták meg egy adatlap kitöltésével. Utánkövetés esetében fontos felmérni további szempontokat is, így a Longitudinális korpuszban egy hosszabb adatlap kitöltését kérjük. Ezzel kívánjuk felmérni, hogy a két felvétel között bekövetkeztek-e esetleg olyan hatások, amelyek miatt a beszédjellemzők nem csak az életkor miatt változhattak.

Az anamnézislapon megjelenő kérdések érintik az adatközlő által ismert beszédproblémáit, azok megjelenési idejét és az ezzel kapcsolatos korábbi beavatkozásokat. Kiterjed az adatközlő hétköznapi hanghasználati szokásaira az ének- és a beszédhang tekintetében is. Rákérdezünk olyan betegségekre (pl. neurológiai vagy mozgásszervi megbetegedések, hormonális jellemzők), illetve a szakirodalomból ismert olyan tényezőkre (pl. alkohol- és folyadékfogyasztás, dohányzás), amelyek hatással lehetnek a beszédre (Hacki, 2013). A kérdőív tartalmaz egy a fonációs zavarok szűréséhez használt eljárásból (Glottal Function Index) kiemelt és kissé módosított kérdéseket, ami így lehetőséget teremt a fonáció aktuális állapotának feltérképezésére (Bach és mtsai, 2005). A fogazat esetleges változásait és a külföldön eltöltött időt is érintik egyes kérdések. Emellett rögzíti az adatközlő súlyát, magasságát és iskolai végzettségét, foglalkozását is. Az anamnézislap előnye, hogy igen részletes, számos területet lefed. Hátránya azonban az, hogy önbevalláson alapul, így csupán azon tényezőkről szerezhetünk információt, amelyet az adatközlő elismer és/vagy bevall.

Mivel nem célzottan, valamely paraméter szerint keresünk beszélőket, hanem a korábbi beszélők közül hívunk vissza adatközlőket, így az egyes tényezők hatását nem tudjuk rendszerszerűen elemezni. Az anamnézislap célja, hogy a beszélők között esetleg megjelenő nagyobb eltérések esetében jó eséllyel kideríthető legyen azok oka.

2.3 Felvételi körülmények

A hangrögzítés körülményei azonosak a BEA és a Longitudinális korpusz felvételein, így ezeket Gósy és munkatársai (2012) alapján foglaljuk össze. Minden esetben a Nyelvtudományi Intézet Fonetikai Osztályán kialakított, zajszigetelt szobában készítjük a felvételeket. A szoba méretei (a hangszigetelő réteget nem számítva): 340×210×300 cm. A hangcsillapítás mértéke a külső környezethez képest 50 Hz-en 35 dB, 250 Hz fölött pedig ≥65 dB. A szoba belső terének fala – az utózengés elkerülése érdekében – 54×54 cm-es hangtörő felületekkel van kialakítva.

A felvételek Audio-technika AT4040 típusú kardioid kondenzátormikrofonnal készülnek. A mikrofon Phonic MM102 típusú többcsatornás phantomtápos analóg keverőn keresztül csatlakozik a számítógéphez. A rögzítés digitális, közvetlenül a számítógépre történik a GoldWave hangeditáló szoftverrel, 44,1 kHz-es mintavételezéssel. (Tárolás: 16 bit, 86 kbyte/s.) A szövegmeghallgatást igénylő feladatok során a visszahallgató típusa Behringer Truth B2031 kétutas aktív stúdiómonitor.

A teljes felvételek hossza az adatközlő beszédkedvének függvénye. A spontán és félspontán beszédfeladatok során, amikor az interjúkészítő úgy látja, hogy az adatközlőnek nincs több mondanivalója, vagy ezt maga az adatközlő jelzi, a következő felvételi egységet kezdik meg.

2.4 A korpusz felvételi egységei

Mivel a korpusz szükségszerűen egy korábbi felvételsorozatból visszahívott adatközlők beszédét tartalmazza, a felvételi egységek egy része azonos, más része közel azonos a korábbi felvételt tartalmazó BEA-adatázis feladataival (Gósy és mtsai, 2012).

A BEA-protokoll nyolc részből áll: mondatismétlés, narratíva/interjú, véleménykifejtés, két tartalomösszegzés, társalgás, mondat- és szövegfelolvasás. Az új utánkövetéses felvételek a Longitudinális korpuszban kilenc részből állnak: mondatismétlés, két narratíva/interjú, véleménykifejtés, két tartalomösszegzés, mondat- és szövegfelolvasás, illetve négyszer visszatérő azonos mondatok felolvasása. Az összetartozó felvételi egységek esetében a Longitudinális korpuszba tartozó felvételen az adatközlőt a korábban, a BEA-felvétel során kapott, érintett témák kifejtésére kéri meg az interjúkészítő. A hanganyagok és a felolvasandó anyagok nagyrészt azonosak. A korábban ismételt és felolvasott 25 mondatból 15-öt válogattunk ki, illetve beépítettünk 17 ismétlődő mondatot, amelyet a Longitudinális-felvétel során négyszer, egymástól független pontokon olvas fel a beszélő.

Mondatismétlés: Az adatközlők feladata ebben a részben a BEA felvételein 25 egyszerű és összetett mondat megismétlése az interjúkészítő után. A mondatok szótagszáma 15 és 26 közé esik. Ha az adatközlő nem emlékszik pontosan az elhangzott mondatra, a felvételvezető megismétli azt. A Longitudinális korpusz felvételein az adatközlőknek ebből a 25 mondatból 15-öt választottunk ki ismétlésre. A feladat ettől eltekintve teljesen azonos a korábbival.

Narratíva/interjú: Ebben a részben az adatközlők feladata, hogy családjukról, munkájukról/tanulmányaikról és hobbijukról beszéljenek. Célja, hogy a beszélő felkészülés nélkül, minél hosszabban beszéljen, az interjúkészítő csak akkor szólal meg, ha úgy ítéli meg, valamilyen kérdés, reagálás továbbviheti a beszélőt. A Longitudinális-felvételek során a 10 évvel korábbi alkalommal feltett kérdésekhez és az ezek mentén esetlegesen felmerült további témákhoz nyúl vissza az interjúkészítő. Azaz, ha egy beszélő főként a munkájáról beszélt, de a családjáról nem, akkor az interjúkészítő ismét a munkájáról kérdezi. A kérdések az elmúlt 10-11 évből indulnak ki, azaz az adatközlőt az azóta ért élményekről kérdezzük.

Véleménykifejtés: Az adatközlő feladata az interjúkészítő által felvetett témák véleményezése. Ha az adatközlő szívesen beszél, az interjúkészítő egy témát vet fel, amennyiben azonban nehezen késztethető hosszabb beszédre egy témában, a felvételvezető további témákkal áll elő. A témák általánosak és hétköznapiak (házasság, illetve együttélés; eutanázia; közlekedés a fővárosban; sztárok/celebek alkohol- és drogproblémái; ittasan okozott autóbalesetek; megúszott büntetések; a magántulajdon védelme; új adók, új szabályozások; sztrájkok; motoros balesetek; szervek felajánlása). Ebben a felvételi egységben gyakrabban szólal meg az interjúkészítő, így a monologikus és dialogikus formák váltakoznak. A Longitudinális felvétel interjúkészítője ismét az előző felvételen elhangzott témában/témákban kérdezi az adatközlőt.

Tartalomösszegzések: Az adatközlő feladata hallott szövegek tartalmának összefoglalása. A protokoll két szöveget tartalmaz: az egyik egy rövid tudománynépszerűsítő cikk (174 szavas; 1 perc 37 másodperc időtartamú), a másik egy történelmi anekdota (270 szavas; 2 perc 5 másodperc időtartamú). Ez a rész majdnem teljesen monologikus, a kísérletvezető nem szólal meg közben, amennyiben nem szükséges ösztönzés az adatközlőnek. A szövegek és a sorrend azonos a BEA- és a Longitudinális-felvételek során.

Társalgás / 2. interjú/narratíva: A BEA-felvételeken három fő társalgását rögzítették, ennek résztvevői az adatközlő, az interjúkészítő és egy további társalgási partner, aki a Fonetikai Osztály egyik munkatársa. A témák a mindennapi élethez kapcsolódnak (pl. karácsony, húsvét ünneplése; mobiltelefon kisgyermekeknek; új KRESZ és a biciklisek; halálbüntetés; dohányzási tilalom a szórakozóhelyeken; éjszakai élet, szórakozási lehetőségek Budapesten). Az adott adatközlő esetében mindig különbözik a véleménykifejtés témájától. Az interjúkészítő a témát az adatközlő érdeklődési körének figyelembevételével választotta ki. A Longitudinális-felvételeken a háromfős társalgást egy interjú-véleménykifejtés váltotta fel. Ebben az adatközlő a korábbi társalgási témát vagy ahhoz hasonlót fejt ki. Ha például korábban a karácsonyi ünnepekről volt szó a BEA-felvétel során, akkor a Longitudinális-felvétel során is családi ünnepekről, kiemelten a karácsonyról szól a fő kérdés.

Felolvasások: Az adatközlő feladata a felvétel elején megismételt 25 (BEA)/15 (Longitudinális) mondat, majd egy tudománynépszerűsítő (291 szavas) cikk felolvasása. Az adatközlőnek lehetősége van előre átolvasnia az adott szöveget.

Visszatérő mondatok: Ez a feladat csak a Longitudinális-korpuszban szerepel. Ez a 17 mondat célzott szegmentális vizsgálatra szolgál, mégpedig felpattanó zárhangok és az *a, e, i, u* magánhangzók elemzésére, amelyek azonos helyzetben jelennek meg a mondatokban több alkalommal. Ezeket a mondatokat négyszer, randomizált sorrendben két-két egyéb felvételi egység között olvassák fel a beszélők.

A legtöbb felvételi egységben tehát hasonló vagy azonos a téma, illetve az alkalmazott szövegek/mondatok. Ennek oka, hogy hasonló hangsorok aktiválódhassanak, így a két felvétel közötti összehasonlításban bizonyos jelenségeket hasonló körülmények között lehessen elemezni.

2.5 Adatkezelés

A felvétel rögzítését megelőzően az adatközlő elolvas egy felvételi leírást, amely tartalmazza a felvétel készítésének okát, a felvételtől való elállási jog ismertetését, illetve egy GDPR alapján készült nyilatkozatot, amely a felvétel kutatási célú felhasználását teszi lehetővé az anonimitás biztosítása mellett. Amennyiben ezen két dokumentumot az adatközlő elfogadja, és aláírásával jelzi, hogy a felvételbe, az adattárolásba és a kutatási felhasználásba beleegyezett, elkészülhet az utánkövetéses felvétel.

A felvételek és az adatlapon adott válaszok tárolása a beszélő nevétől függetlenítve történik. az új felvétel kódja alapján a korábbi felvétel egyértelműen visszanyerhető, azonban az sem tárolódik együtt az adatközlő nevével.

2.6 Hanganyagok, annotációk

A BEA-felvételek lejegyzése eddig Wordben, Transcriberben és Praat szoftverrel (Boersma–Weenink, 2019) történt (Gyarmathy és mtsai, 2014, Neubgerger és mtsai, 2014). A Longitudinális korpusz esetében a lejegyzéseket a BEA praatos lejegyzéseire alapozzuk, mivel ezek a TextGrid-fájlok több szoftverben is alkalmazhatóak, könnyen

átalakíthatóak, és időhöz rendeltek. Az alábbiakban összefoglaljuk Gyarmathy és munkatársai (2014) és Neuberger és kollégái (2014) tanulmányai alapján a praatos lejegyzés főbb jellemzőit, melyeket a két korpusz (BEA és a Longitudinális is) követ.

A lejegyzés háromszintű: beszédszakasz-, szó- és beszédhangszintet tartalmaz. Ez a beszélők számától függően 6 (2 beszélő esetén), illetve 9 (3 beszélő esetén) címkesort jelent. A lejegyzések során a címkesorok sorrendje és elnevezése állandó.

A beszédszakaszszintű lejegyzések helyesírásban (fonémaalapon), nagybetű használata és központozás nélkül készültek. Egy beszédszakasz a beszélő által tartott szünettől szünetig tartó egység (a határoló szünet lehet néma vagy kitöltött). A szószintű lejegyzés szintén helyesírásban (fonémaalapon) történik. A beszédhangszintű lejegyzés (az előző két szinttől eltérően) már nem fonéma-, hanem beszédhangalapon történik. Egy beszédhangot minden esetben egy karakter jelöl (a kettős betűket nagybetűk helyettesítik). A nyelvileg hosszú mássalhangzókat a mássalhangzó betűjének egyszeri leírása utáni kettőspont jelzi.

Nagybetűkkel történik a beszédrészeken felül megjelenő egyéb nyelvi elemek jelölése, szintén egységes jelölésrendszerben (pl. SIL – néma szünet, Ö, M, ÖM – hezitálás, KUKA – zaj miatt elemzésre használhatatlan beszédrészek).

A nevetést mindhárom szinten jelölik, ha az adott egység valamely részét a beszélő nevetve mondja, akkor ezt az egység elején jelzi a NEV jelölés. Az egyszerre beszélések szintén mindhárom szinten jelölve vannak, ahol két vagy három beszélő is beszél, ott a szöveg nem jelenik meg a lejegyzésben, hanem az EB jelölés jelzi az egyszerre beszélést. Ha a beszélő nem szótári alakban előforduló szót ejt (pl. nyelvbotlás, egyszerűsítés: *nemtom, asszem, szal* stb.), akkor beszédszakaszszinten az elhangzott alakot, illetve utána szögletes zárójelben a szótári alakot (pl. *asszem [azt hiszem])* találjuk meg. Szószinten és beszédhangszinten kizárólag az ejtett alakokat tartalmazzák a címkék. A kérdések jelölése beszédszakaszszinten történik, mivel az írásjelek használata a lejegyzések során mellőzött, ezért a beszédszakasz elején megjelenő Q jelzi a kérdést.

A helyesírás szerinti kötőjelek nem részei a lejegyzésnek, a kötőjeles szavakat egybeírjuk (pl. *spontánbeszédadatbázis*). A számok betűvel kiírva, a helyesírásnak ellentmondóan kötőjel nélkül egybeírva szerepelnek (pl. *kétezerháromszáztízben*). A kötőjelek a szótöredékek jelzésére szolgálnak (pl. *palacsin-*). A betűszavak lejegyzése kisbetűvel történik (pl. *eltére*), beszédszakaszszinten feloldásra kerülnek, szintén szögletes zárójelek között (pl. *[ELTEre]*). Az idegen szavak a kiejtett formájukban vannak lejegyezve (pl. *kvescsön*), illetve beszédszakaszszinten ez esetben is megtörténik a feloldás szögletes zárójelek között (pl. *[question]*).

Az eredeti lejegyzési rendszert kiegészítve a mondatismétlés, tartalomösszegzés és felolvasás esetében egy további címkesor vált szükségessé, mivel ezeknél a beszédtípusoknál előfordulhat, hogy az adatközlő megjegyzéseket fűz saját magához, tehát például a felolvasás részben előfordulhatnak spontán beszédrészek is. Ezen részek kiszűrése az egységesen *Komment* névvel feltüntetett címkesorokkal történik.

3 Felhasználási lehetőségek

A jelen tanulmányban egy folyamatban lévő, felnőttek beszédét 10-11 év távlatában rögzítő longitudinális korpuszt mutattunk be. Szándékunk a korpuszt más kutatók számára is hozzáférhetővé tenni. A beszédkorpusz a lejegyzésével számos vizsgálathoz

evidensen hozzájárulhat. Utánkövetéses mivolta révén pedig a korábbi longitudinális vizsgálatok viszonylag kisszámú adatközlőn kapott eredményeiből felmerült kérdésekre adhat választ. Egyrészt szükségszerűen eltérő eredmények születhettek egy-egy beszélő ejtésében, másrészt a speciális adatközlői csoportok olyan tényezőket is magukkal vontak, amelyek az (elsődleges és másodlagos) életkori változáson és az esetleges diakrón hatáson túl is befolyásolták az eredményeket. Mindemiatt a nagyobb adatközlőszámon végezthető vizsgálatok tovább árnyalhatják az eredményeket. Az utánkövetéses vizsgálatok ugyanakkor nem csak a fonetikai, de további vizsgálatok elvégzéséhez is hozzájárulhatnak, így pl. a beszélő azonosításához.

Köszönetnyilvánítás

A kutatást a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal NKFIH-FK-128814 számú pályázata támogatta.

Hivatkozások

Ash, S., Jester, C., York, C., Kofman, O. L., Langey, R., Halpin, A., Firn, K., Perez, S. D., Chahine, L., Spindler, M., Dahodwala, N., Irwin, D. J., McMillan, C., Weintraub, D., Grossman, M.: Longitudinal decline in speech production in Parkinson's disease spectrum disorders. Brain Lang 171, 42–51. (2017)

Bach, K. K., Belafsky, P. C., Wasylik, K., Postma, G. N. & Koufman, J. A.: Validity and reliability of the Glottal Function Index. Archives of Otolaryngology - Head & Neck Surgery 131(11), 61–64. (2005)

Blair, M., Marczinski, CA., Davis-Faroque, N. & Kertesz, A.: A longitudinal study of language decline in Alzheimer's disease and frontotemporal dementia. J Int Neuropsychol Soc. 13(2). 237–245 (2007)

Boersma, P., Weenink, D.: Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/download_win.html. (A letöltés ideje: 2019. október 1.), (2019)

Busse, E. W.: General Theories of Aging. In: Copeland, J. R. M., Abou-Saleh, M. T., Blazer, D. G. (szerk.) Principles and Practice of Geriatric Psychiatry, Second Edition. Willey-Blackwell (2002)

Decoster, W., Debruyne, F.: Longitudinal Voice Changes: Facts and Interpretation. Journal of Voice 14(2), 184–193 (2000)

Gósy, M.: Long-term within-speaker and between-speaker differences in phonetic output: Evidence from Hungarian. In: Braun, A., Masthoff, H. R. (szerk.) Phonetics and its Applications. Festschrift for Jens-Peter Köster on the Occasion of his 60th Birthday. pp. 75–85. Steiner, Stuttgart (2002)

Gósy M., Gyarmathy D., Horváth V., Gráczi T. E., Beke A., Neuberger T., Nikléczy P.: BEA: Beszélt nyelvi adatbázis. In: Gósy M. (szerk.) Beszéd, adatbázis, kutatások. pp. 9–24. Akadémiai Kiadó, Budapest (2012)

Gósy M., Krepsz V.: Magánhangzók temporális jellemzői az idő múlásának függvényében. Beszédkutatás 23, 53–65 (2015)

Gráczi T. E., Krepsz V.: Évek múltán a zönge: Egyes zöngejellemzők változása 11 év alatt 6 férfi beszélő beszédében. MANYE 2019. évi kongresszusának proceedingskötete. (meg. alatt)

Guyuron, B., Rowe, D. J., Weinfeld, A. B., Eshraghi, Y., Fathi, A., Iamphongsai, S.: Factors contributing to the facial aging of identical twins. Plast Reconstr Surg. 123(4), pp. 1321–1331 (2009)

Gyarmathy D., Neuberger T., Gráczi T. E.: Lejegyzési útmutató a BEA Spontánbeszéd-adatbázis háromszintű annotálásához. Alkalmazott Nyelvtudomány 14(1), 35–44 (2014)

Hacki T.: A beszéd- és énekhangképzés fiziológiája, akusztikája, patológiája és terápiája. In: Hirschberg J., Hacki T., Mészáros K. (szerk.): Foniátria és társtudományok I. pp. 85–272. ELTE Eötvös Kiadó, Budapest (2013)

Harrington, J., Palethorpe, S., Watson, C. I.: Does the Queen speak the Queen's English? Nature 408. 927. (2000a)

Harrington, J., Palethorpe, S., Watson, C. I.: Monophthongal vowel changes in Received Pronunciation: an acoustic analysis of the Queen's Christmas broadcasts. Journal of the International Phonetic Association 30(1/2), 63–78 (2000b)

Harrington, J., Palethorpe, S. & Watson, C. I.: Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In Proceedings of Interspeech 2007, pp. 2753–2756. (2007)

Hunter, E. J., Kapsner-Smith, M., Pead, P., Engar, M.Z., Brown, W. R.: Age and speech production: a 50-year longitudinal study. Journal of the American Geriatrics Society 60(6), 1175–1177 (2012)

Jacob, M. E., Ganguli, M.: Epidemiology for the clinical neurologist. In: Rosano, C., Ikram, M. A., Ganguli, M. (szerk.) Handbook of Clinical Neurology 136, 3–16. (2016)

Labov, W.: Principles of Linguistic Change, Internal Factors. Wiley-Blackwell, Oxford (1994)

Labov, W.: Principles of Linguistic Change, Volume 2. External Factors. Blackwell, Oxford (2001)

Lalley, P. M.: The aging respiratory system-Pulmonary structure, function and neural control. Respiratory Physiology & Neurobiology 187, 199–210 (2013)

Le, X., Lancashire, I., Hirst, G., Jokel, R.: Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. Literary and Linguistic Computing 26(4), 435–461 (2011)

Liu, L. F., Su, P. F.: What factors influence healthy aging? A person-centered approach among older adults in Taiwan. Geriatrics & Gerotology International 17(5), 697–707 (2017)

Misono, S., Marmor, S., Roy, N., Mau, T., Cohen, S. M.: Multi-institutional Study of Voice Disorders and Voice Therapy Referral: Report from the CHEER Network. Otolaryngol Head Neck Surgery 155(1), 33–41 (2016)

Neuberger T., Gyarmathy D., Gráczi T. E., Horváth V., Gósy M., Beke A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: Proceeding of the 17th International Conference, TSD 2014, September 8-12, 2014., Brno. pp. 424–431. (2014)

Powell, M., Filter, M. D., Williams, B.: A longitudinal study of the prevalence of voice disorders in children from a rural school division. Journal of Communication Disorders 22(5), 375–382 (1989)

Quené, H.: Longitudinal trends in speech tempo: The case of Queen Beatrix. The Journal of the Acoustical Society of America 133, EL452.. 452–457 (2013)

Reubold, U., Harrington, J., Kleber, F.: Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. Speech Communication 52, 638–651 (2010)

Russell A., Penny L., Pemberton C.: Speaking fundamental frequency changes over time in women: a longitudinal study. Journal of Speech Language and Hearing Research 38, 101–109. (1995)

Trudgill, P. J.: Introduction: Sociolinguistics and sociolinguistics. In: Trudgill, P. J. (szerk.) Sociolinguistic Patterns in British English London: Edward Arnold. 1–18. (1978)

Verdonck-de Leeuw, I. M., Mahieu, H. F.: Vocal aging and the impact on daily life: a longitudinal study. Journal of Voice 18(2), 193–202 (2004)

Szaknyelvi annotációk javításának statisztikai alapú támogatása

Kicsi András[1], Pusztai Péter[1,2], Szabó Endre[3], Vidács László[1,2]

[1]Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
Szeged, Dugonics tér 13.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
[3]Szegedi Tudományegyetem
Szeged, Dugonics tér 13.
{akicsi,pusztaip,lac}@inf.u-szeged.hu, endrebacsi@gmail.com

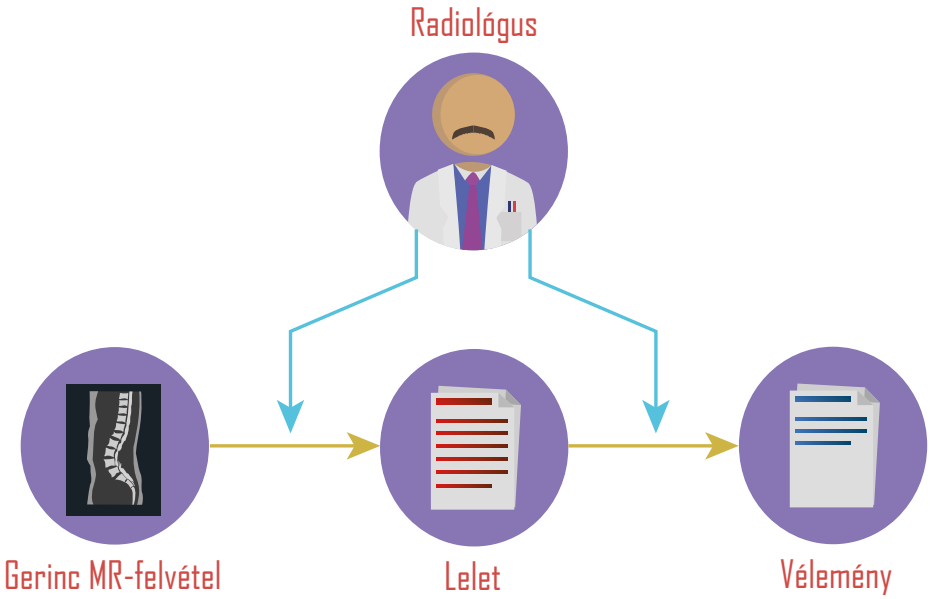
Kivonat A radiológiai leletezés komoly feladat, melynek automatizálása nagy jelentőséggel bír. A leletek gépi értelmezéséhez tanítópéldákra van szükség, amelyeknek megfelelő minőségben kell előállnia. Jelen munkában egy olyan módszert mutatunk be, amellyel az annotáció konzisztenciájának javítása érdekében, az újbóli átnézést statisztikai módszerekkel támogattuk, az inkonzisztenciákra az annotációs rendszer felületén hívva fel a figyelmet. Módszerünk eredményességét valós eredményekkel támasztjuk alá, amelyek nem csak a konzisztenciára, hanem a gépi tanulás sikerére is nagy mértékben kihatnak.
Kulcsszavak: radiológia, információkinyerés, nlp, annotáció

1. Motiváció

A klinikai leletezés az orvoslás jelentős területe, amelynek sikere nagyban hozzájárul a páciensek végső gyógyulásához. Ezen belül a radiológia területén végzett gerinc MR vizsgálatok is igen gyakoriak, csak Magyarországon évente sok ezer ilyen lelet készül. Ezeket általában természetes nyelvű szöveggel, magyar nyelven fogalmazzák meg a radiológusok. A vizsgálaton készített képeket szemlélve leírják orvosi szakértelmüknek megfelelően a látott elváltozásokat, így készülnek el a leletek, és a hozzájuk tartozó, tömörebb vélemények. Ezt a folyamatot láthatjuk az 1. ábrán.

A leletezés természetesen nem könnyű feladat, és folytonos odafigyelést igényel az orvos részéről. Ez azonban könnyíthető különböző automatizáló megoldásokkal, mint például lehetőség a leletek diktálására gépelés helyett. Amennyiben a leleteket automatizált módon értelmezni is tudnánk, rengeteg egyéb lehetőség nyílna meg a munka segítésére. Ezek felhasználhatók lennének mind a minőség biztosításában, mind a leletezés zökkenőmentesebbé és gyorsabbá tételében. Kutatásunkkal ezt tűztük ki célul, melyhez első lépésként a szövegben előforduló entitások detektálását tekintjük. Korábbi munkánkban (Kicsi és mtsai, 2019) már

publikáltuk a területen végzett annotációs módszerünket, illetve kezdeti eredményeinket is. Módszerünkben testrészeket, elváltozásokat és tulajdonságokat különböztettünk meg a leletek szövegében, melyeket automatikusan detektáltunk. Testrésznek tekintettük az emberi test egy pontosan megnevezett elemét, mint például „L.V. discus”, elváltozás minden kóros eltérés, mint például „előboltosulás”, de az aspektusok, például „magassága” és pozitív állapotot jelző szavak, mint például „ép” is ide tartoznak. Tulajdonság minden olyan mértéket vagy minőséget leíró kifejezés, amely elváltozást pontosít, mint például „3 mm-es” vagy „körkörös”. A megfelelő detektáláshoz tanulóadatokra van szükség, ezeket egy radiológus segítségével annotáltattuk, melyhez a Brat (Stenetorp és mtsai, 2012) annotációs szoftvert használtuk fel.



1. ábra: A radiológus munkája a vizsgálat után

Kezdeti detektálási kísérleteink után hamar nyilvánvalóvá vált, hogy a meglévő 487 lelet annotációja jelen minőségben nem elég valóban kiváló eredmények előállításához. Természetesen erre egy lehetséges módszer másik radiológus alkalmazása és a két annotáció összehasonlítása, mely munka azóta szintén megtörtént, ám kezdeti annotációink minőségét is javítani kívántuk, mivel számos inkonzisztenciát tapasztaltunk a jelölésekben. Noha az annotációs útmutatót igyekeztünk pontosan előállítani, mégis felmerült nagy mennyiségű egyéni döntés és dilemmás eset, amelyen a radiológus gyakran önmagával sem tudott konzisztens maradni.

Egy újbóli annotáció természetesen igen nagy feladat még akkor is, ha csak a hibás eseteket kell kijavítani. Arra sincs semmi garancia, hogy ezúttal fenntartható a folyamatos konzisztencia. Ezért automatizált módszerrel igyekeztünk ezt elősegíteni. Cikkünkben az erre kifejlesztett statisztikai módszerünket mutatjuk be, amely Brat rendszer által kimenetként adott .ann fájlok vizsgálata után tokenenként állapít meg konzisztenciát, az eredményeket pedig a Brat formátumának megfelelően rögzíti megjegyzésként. Ezzel felhasználóbarát módon hívja fel a figyelmet a kevéssé konzisztens jelölésekre, amelyek tudatában a radiológus ezután teljes mértékben saját elbírálása szerint járhat el.

Korábbi cikkünkben említettük, hogy testrészek, elváltozások és tulajdonságok mellett helyeket is jelöltünk, ezek a munka jelenlegi fázisában azonban komplex szerkezetűek, így velük itt nem foglalkozunk.

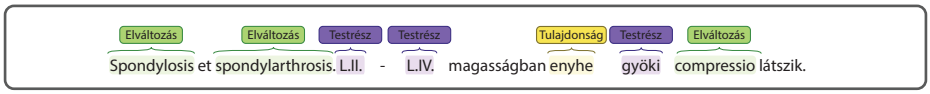
2. Folyamat

Munkánk jelenleg magyar nyelvű gerincleletek feldolgozását öleli fel. Cikkünkben a helyes klasszifikáció problémájával foglalkozunk, amelyben a leletek szövegének kifejezéseit három címkével igyekszünk ellátni jelentésüknek megfelelően, testrészeket, elváltozásokat és tulajdonságokat különböztetünk meg. A problémát gépi tanulási módszerekkel közelítettük meg. Mindkét módszer nagy mennyiségű tanulóadattal tud csak megfelelően működni, ezért ezt biztosítani kell. Erre a célra radiológus által annotált valós leleteket használtunk, 487 lelet annotációja készült el. Ehhez radiológusunk a Brat (Stenetorp és mtsai, 2012) annotációs rendszert használta, amelyet megfelelően konfiguráltunk a kívánt entitások jelölésére, így áttekinthető és felhasználóbarát környezetben végezhette a jelölést.

Ezen az annotált leletmennyiségen igyekeztünk javítani egy statisztikával támogatott újabb kézi elbírálással. Az annotációs útmutató számos esetet lefed, ám ezeket többszáz lelet átolvasása után már nem mindig idézi fel az annotátor helyesen. Vannak továbbá olyan különleges esetek, amelyek egyszerűen nem illenek semelyik, az útmutató által érintett problémakörbe. Ez utóbbiakat jobb esetben megbeszélés alapján kell kezelni, ám sok olyan eset adódik, amikor az annotátor eléggé biztosnak tart egy bizonyos helyzetet, és önállóan jelöli. Ilyenkor a legfontosabb, hogy önmagával konzisztens legyen a felmerülő hasonló dilemmás kérdéseket mindig egy irányelv mentén jelölje.

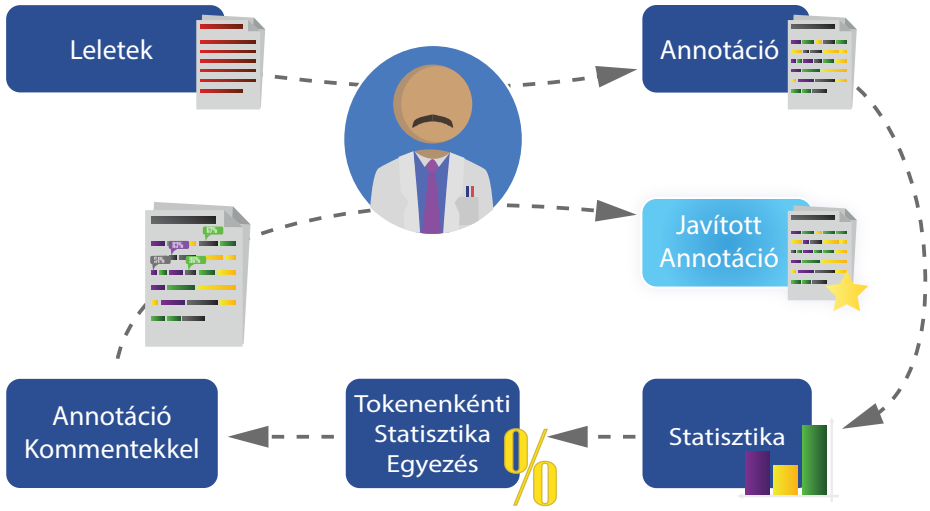
Az annotátor önmagától való inkonzisztenciája egyszerű statisztikai módszerekkel könnyen mérhető, megtekinthetjük, hogy egy adott kifejezést általában ugyanazon címkével látta-e el. Ez persze rengeteg esetben indokolt kilengés, mint például munkánkban a „jobb” szó esetében, ahol ez lehet tulajdonság része, egy testrész helyének pontosítása, vagy akár annak leírása, hogy egyik csigolya a másiknál jobb állapotban van, amely elváltozás lenne. Tehát az emberi elbírálás semmiképpen sem nélkülözhető.

A statisztikák azonban segíthetnek a kézi ellenőrzésben, nagyban felgyorsíthatják azt, és felhívhatják figyelmet olyan dilemmás esetekre, amikre az emberi személő esetleg nem is figyelt volna fel. Tekintsük a 2. ábrán látható példát. Itt különböző dilemmás esetek merülnek fel. Először is a „spondylosis et spondyl-



2. ábra: Egy több dilemmás esetet tartalmazó példa

arthrosis" szövegrész nagyon sokszor ugyanígy fordul elő a leletekben, ugyanis leggyakrabban a két elváltozás együtt jelentkezik. Ez kísértést jelent az annotátor számára, hogy egyben jelölje őket. Az „et" szó továbbá, mivel latinul van, kevésbé intuitívan tagol elváltozásokat, mint például az „és" szó. Másrészt láthatjuk, hogy ahogy a leletek többségében, itt is vannak intervallummal megadott testrészek. Ilyenkor megegyezés és az annotációs útmutató szerint ezeket külön-külön be kell jelölni. Efölött azonban könnyű átsiklani, hiszen tömörek, egyértelműen testrészt jelölnek, és szinte teljesen egyben vannak, még szóköz sincs közöttük az eredeti szövegben. Ezért rengeteg hasonló típusú hiba volt a kezdeti annotációban, amely a testrészek inkonzisztens detektálását idézte elő egyes esetekben. A „gyöki compressio" kifejezést is gyakran egyben jelölte a radiológus, hiszen úgy tűnik, hogy ez az elváltozás teljes megnevezése. A „gyök" azonban egy testrész és más megfogalmazásban így is jelölné a radiológus is, már igen apró változtatás után is, például „a kilépő gyökök compressioja látható" formában. Az ilyen típusú hibák szintén nagyon gyakoriak.

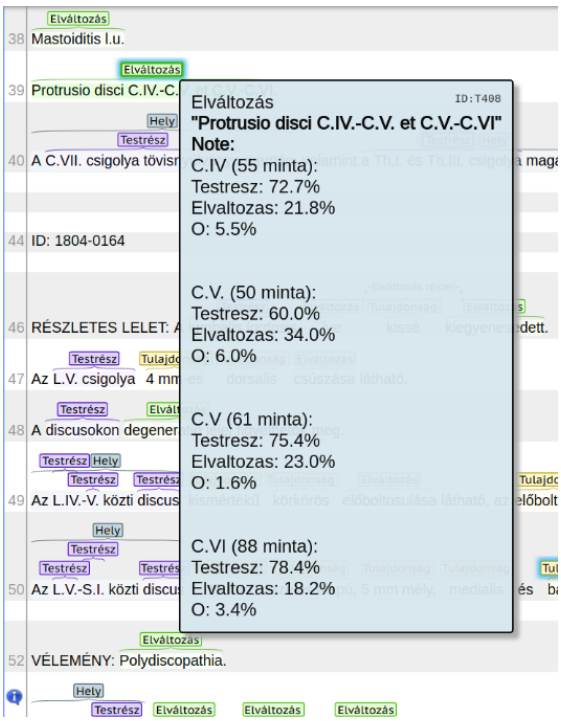


3. ábra: A javasolt javítási módszer áttekintése

Természetesen ezek nagy odafigyeléssel kijavíthatók egy újbóli átnézés során, de ennél sokkal jobb módszer lehet, ha megpróbálunk ezekre automatikusan rávilágítani. Az erre kidolgozott módszerünk látható a 3. ábrán. A 487 lelet összes szavát tokenek szintjén listába gyűjtöttük. Ezután minden egyes tokenhez meghatároztuk, hogy hány esetben voltak testrészként, elváltozásként és tulajdonságként jelölve, tehát előállítottuk a szükséges statisztikát. Ennek a listának a birtokában az összes leleten végigiterálva és tokenizálva minden egyes token előforduláshoz meghatároztuk, hogy a többségi címkéjükben vannak-e, három címke esetén is a legnagyobbhoz viszonyítottunk. Egyenlőség esetén mindkét címkét többséginek vettük. Amennyiben egy token nem a többségi címkéje részeként szerepelt egy jelölésben, akkor ezt a Brat rendszer által kimenetként adott .ann fájlban jelöltük. Az annotációs rendszer megengedi megjegyzések beszúrását egyes címkézett elemekhez. Ilyenkor praktikus módon a címke fejléce ragyogó körvonalat kap, szembetűnően felhívva magára a figyelmet. A fájlba tehát a Brat rendszer formátumával teljesen megegyező új sorokat szúrtunk be automatizáltan, amelyben leírtuk, hogy a token nem a többségi címkéjében fordult elő. Egy jelölt kifejezéshez több ilyen megjegyzés is tartozhat, hiszen sokszor több token szerepel egymás mellett, ilyen esetekben az összes megjegyzés egymás alá kerül. Miután az algoritmus az összes leletet átnézte, és előálltak a módosított .ann fájlok, a Brat rendszerrel megnyitva a leleteket már láthatjuk a kék színnel ragyogó címkéket, amelyek remekül felhívják a figyelmet a dilemmás helyzetekre. Erre látható példa a 4. ábrán, amely egy képernyőkép módszerünk kimenetéről. További előnyt jelent, hogy a radiológusnak sem kell új rendszerrel megismerkedni, a már megszokott környezetben végezheti a leletek átvizsgálását.

Az ábrán látható példán az annotátor a diagnózis egy teljes mondatát egy elváltozásként jelölte, a rendszer pedig felhívja a figyelmet arra, hogy a csigolyák megnevezése általában testrész szokott lenni. A számok nyomán egyébként gyanítható lenne, hogy máskor is csinált már ilyet. Az "O" címke azt jelöli, hogy nem volt jelölve egyik címkével sem. A 2. ábrán látható példában a megnevezett rossz jelölések esetén módszerünk ugyanígy szólna például, hogy az „et” szó és a kötőjel általában nem szokott jelölésre kerülni ha külön tokenként fordul elő, a „gyöki” szó pedig általában testrész.

Biztosítottunk továbbá egy megértést segítő modult is a statisztikák mellett. Ez egy egyszerű programkód, amely szöveget vár bemenetként, és az összes leletet végigpásztázva megadja, hogy milyen címkével, hol, és milyen szövegkörnyezetben volt jelölve az adott kifejezés. Ez azokra az esetekre alkalmazható, ha esetleg nem értjük, hogy az adott szó mi alapján került egy kisebbségi címkébe, hiszen jelenleg jó jelölést látunk rá. Ezután akár a többi ilyen eset, vagy esetleg hibásan kiosztott többségi címke is könnyen megkereshető és javítható. A statisztikát kiszámító és kommenteket beszúró programkód is a radiológus rendelkezésére állt, amelyet bármikor újrafuttathatott, hiszen ezek nem aktualizálják magukat automatikusan.



4. ábra: Képernyőkép kimenetünk Brat-ban való megjelenítéséről

3. Kísérletek

Kísérleteink célja olyan eszköz fejlesztése volt, mellyel segíteni tudjuk a radiológust a tekintetben, hogy az általa készített annotáció egyrészt önmagával, másrészt az annotálási útmutatóval is minél inkább konzisztens maradjon. Első lépésben 487 gerinc MR leletet annotáltattunk a radiológus kollégával az előre meghatározott útmutató szerint. Már a folyamat közben is kimutatható volt, hogy az annotáció nem teljesen konzisztens, amit a radiológus kolléga is alátámasztott, azzal a megfigyelésével, hogy sokszor nem hogy az útmutatóval, de még önmagával sem tudja tartani a konzisztenciát egy hosszabb annotálás során. Az annotációban tapasztalható inkozisztenciák felméréséhez minden egyedi tokenhez statisztikát készítettünk, melyben kimutattuk a különböző címkék tokenhez rendelésének százalékos eloszlását. A 487 leletben 2760 egyedi tokent találtunk, melyből 2082 tokenhez kizárólag egyféle címke lett rendelve. A maradék 678 tokent a radiológus minimum kétféleképpen annotálta. Az inkonzisztencia sok esetben adódott a kifejezés különböző szövegkörnyezetben történő előfordulásaiból, aminek következtében a radiológus eltért az útmutatótól és saját legjobb belátása szerint annotált. Ezekhez az annotációs esetekhez, mivel statisztikailag gyakran kisebbségben voltak, egy figyelmeztető megjegyzést rendeltünk, melyet az annotáló szoftverben jelenítünk meg. Ezek alapján a radiológus belátása sze-

rint döntött az eset javítása, vagy változatlanul hagyása mellett, természetesen az annotációs útmutató által lefektetett alapelvekkel továbbra is összhangban maradva.

Az általunk fejlesztett segédeszközökkel felvértezett radiológust ezután egy javítóannotációra kértük. A visszakapott leletekben a 2760 egyedi tokenből ezután már 2276 token rendelkezet kizárólag egyféle annotációval a többi minimum kétféle címkét kapott. Már ez a szám is mutatja, hogy a korábbi annotációhoz képest konzisztensebb eredményt kaptunk a javítást követően. A többféle címkével annotált tokenek pontos eloszlását az első és második körös annotáció után az 1. táblázat szemlélteti. A táblázatban is jól látható az annotációk javításának eredményessége. Javítás után a három- és négyféle címkét kapott tokenek mennyisége a felére, míg a kétféle címkét kapott tokenek mennyisége az eredeti ötödével csökkent.

1. táblázat. A többféle annotációt kapott tokenek eloszlása annotációjavítás előtt és után.

Annotáció	Eredeti	Javított
Egyféle	2080	2276
Kétféle	493	392
Háromféle	156	77
Négyféle	31	15
Összesen	2760	2760

Következő lépésben intra-annotátor egyezést mértünk a radiológusunk eredeti és javított annotációja között. Az egyezés minőségének megítéléséhez a Cohen kappa mutatót, valamint mikroátlag F1-mérték metrikákat alkalmaztunk, ahol a referenciának a javított annotációt vettük. Cohen kappára 0,9278-as értéket, míg F1-mértékre 0,9350-es értéket kaptunk. A szakirodalom szerint a 0,8-as érték feletti Cohen kappa jó egyezésre utal, valamint a magas F1-mérték is azt sugallja, hogy az eredeti és javított annotáció egymással konzisztensnek számít.

A fentiek alapján azt gondolhatnánk, hogy az annotációjavításnak nem sok jelentősége volt, azonban érdekesebb eredményeket kapunk, ha megvizsgáljuk az eredeti és javított annotációval tanított modellek teszthalmazon mutatott teljesítményét. Demonstrációs céllal, a kísérleteinkben referenciaként használt, IOB címkéket nem tartalmazó osztálycímkékkel tanított Bi-LSTM (Hochreiter és Schmidhuber, 1997) eredményeit mutatjuk be a 2 és 3. táblázatokon.

A tanítási eredmények jól mutatják, hogy az annotációk javítása jelentős mértékben javított a modellünk teljesítményén. A testrész és tulajdonság esetében több, mint 3%-os, míg az elváltozás tekintetében megközelítőleg 2%-os javulást értünk el az F1-mértéket tekintve. Érdemes kiemelni, hogy a testrészek felismerési pontossága majdnem 5%-kal nőtt, ez is tükrözi, mennyire jellemző volt az annotációban a testrészek fent bemutatott tipikus inkonzisztens jelölése.

2. táblázat. Az eredeti annotáción tanított Bi-LSTM modell teljesítménye a három fő névelemtípus felismerésében.

	Pontosság	Fedés	F1-mérték
Elváltozás	0,9143	0,9285	0,9213
Testrész	0,9112	0,9519	0,9311
Tulajdonság	0,8856	0,8610	0,8731
Mikroátlag	0,9083	0,9253	0,9167

3. táblázat. A javított annotáción tanított Bi-LSTM modell teljesítménye a három fő névelemtípus felismerésében.

	Pontosság	Fedés	F1-mérték
Elváltozás	0,9380	0,9432	0,9406
Testrész	0,9598	0,9698	0,9648
Tulajdonság	0,9108	0,9020	0,9064
Mikroátlag	0,9420	0,9464	0,9442

Kísérleteink jól szemléltetik, hogy az elsőre kiemelkedően jónak tűnő intra-annotátor egyezés megtévesztő lehet, a számok mögé tekintve, a modelleket a tényleges adatokon tesztelve láthatjuk, hogy az annotáció konzisztenciájának javítása jelentős javulást eredményezhet a modellek működését illetően.

4. Kapcsolódó kutatások

Ugyan próbálkozások történtek már a strukturált leletezés egészségügyi szektorba történő bevezetésére, a gyakorlat napjainkig azt mutatja, hogy a szakorvosok és radiológusok is előnyben részesítik a szabad megfogalmazású leletek készítését a strukturált leletezéssel szemben. Ez egyfelől lehetőséget ad a természetes nyelv komplexitásának kihasználására és a leletek szabatos megfogalmazásra, másfelől megnehezíti a leletekből történő információkinyerést, szövegértelmezést, illetve a leletezési folyamat minőségbiztosítását. Éppen ennek a kihívásnak köszönhetően a terület kiváló kutatási lehetőséget biztosít a számítógépes nyelvészet számára, mely során újfajta természetesnyelv feldolgozási módszerek, illetve az egészségügyi szakembereket segítő alkalmazások egyaránt napvilágot láthatnak.

Az eddig fejlesztett alkalmazások köre a kinyert információ típusától függően széles spektrumon változik. Többek között beszélhetünk diagnózissegéd (Pham és mtsai, 2014; Rink és mtsai, 2013; Solti és mtsai, 2009), diagnosztikai minőségbiztosítást (Raja és mtsai, 2012; Ip és mtsai, 2011; Sistrom és mtsai, 2009; Dang és mtsai, 2008), a leletek automatikus BNO kódolását végző (Farkas és Szarvas, 2007), a nem várt elváltozásokra adott válaszlépéseket (Dutta és mtsai, 2013), vagy a további vizsgálatokra vonatkozó ajánlásokat figyelő (Yetisgen-Yildiz és mtsai, 2011), illetve a páciens egészségi állapotát nyomon követő al-

kalmazásokról (Cheng és mtsai, 2010). A közelmúltban több olyan összefoglaló cikk is megjelent, mely jól bemutatja az elmúlt egy évtizedben történt fontosabb előrelépéseket (Wang és mtsai, 2018; Pons és mtsai, 2016; Ford és mtsai, 2016; Cai és mtsai, 2016; Yim és mtsai, 2016; Meystre és mtsai, 2008).

A leletekből történő információkinyerés első lépése továbbra is a szöveg, előre meghatározott útmutató alapján, szakember által végzett, pontos annotálása. Az annotálást minden esetben minimum két annotátor egymástól függetlenül végzi. A nem egyértelmű esetek eldöntése kettőnél több annotátor esetében többségi szavazással történik, míg két annotátor esetében vagy megegyezéses alapon, vagy egy harmadik, szenior kolléga döntése alapján oldják fel az ellentétet. Az egyezés mérésére, az annotátorok számának függvényében többféle metrikát is alkalmaznak, azonban az egyik legelterjedtebb ilyen mérőszám a Cohen kappa (Artstein és Poesio, 2008). A mérőszám interpretálása a szakirodalomban vita tárgyát képezi. Általánosságban elmondhatjuk, hogy 0,8-as érték felett az egyezés megalapozottnak, az annotáció minősége pedig jónak mondható, ennek hitelességét azonban egyes kutatók megkérdőjelezik (Klebanov és Beigman, 2009). Szerintük ugyanis a magas kappa érték főleg két annotátor esetében nem feltétlen jelent jó minőségű annotációt, csakúgy, mint ahogy az alacsony kappa érték, öt annotátor esetében nem feltétlen jelent rossz minőségű annotációt.

Egy modell maximum annyira lehet jó, mint amennyire jó az adat, amin tanították. Az annotáció minőségének javítása ezért komoly kihívás a számítógépes nyelvészek számára. Ennek elérése érdekében több megközelítést is alkalmaznak. Az egyik legkézenfekvőbb módszer az adat előannotálása, majd az automatikusan létrehozott annotációk szakemberrel történő hitelesítése, illetve javítása. Az előannotáció történhet egy már meglévő adatbázis, vagy az annotátor korábbi annotációja alaján (pl. az annotátor annotálja az adatok felét, majd ezek alapján megtörténik az adatok másik felének automatikus annotációja, amit az annotátor jóváhagy, vagy javít (Ganchev és mtsai, 2007)). Ilyen támogatás több annotáló szoftverben is megtalálható. Egy másik lehetőség az annotációk minőségének javítására, ha annotáció közben ajánlásokat teszünk az annotátornak. Ez annyiból kifinomoltabb, mint az előannotálás, hogy ebben az esetben a korábban többféleképpen annotált esetekre egy megbízhatósági értéket is biztosítunk. Vagyis minden egyes szóra az ajánlást az adott szóhoz korábban hozzárendelt osztálycímkék háttérben kiértékelt statisztikája alapján hozzuk létre. Az ajánlás tehát több osztálycímkét is tartalmaz egy százalékos megbízhatósági érték kíséretében (Oliveira és mtsai, 2017; Morton és LaCivita, 2003). Az MIT fejlesztése a Story Workbench szoftver, mely automatikus annotálási funkcióval is el van látva. Ez annyiban különbözik az előannotálástól, hogy itt az annotációk az annotálás során, a módosításokat figyelembe véve, valós időben keletkeznek (Finlayson, 2011). A WebAnno egy másik félautomata annotációs eszköz, melyben az annotációs javaslatot egy külön ablakban jelenítik meg, az éles szövegen csak a már elfogadott javaslatok, illetve az annotátor által kézzel készített annotációk láthatóak. Ez a konstrukció a szerzők szerint arra ösztönzi az annotátort, hogy minden egyes javaslatot jóváhagyjon, mielőtt az az éles szövegbe kerülne. A program egyébként többrétegű ajánlási rendszert alkalmaz, melynek egyik rétege egy

adott szó korábbi annotációinak későbbi esetekhez rendelése egyszerű szöveges egyezés alapján (Muhie és mtsai, 2014). A GoNTogle egy szemantikus annotációt ellátó eszköz, mely teljes dokumentumok vagy dokumentum részek automatikus annotálására is képes. Az automatikus annotációhoz egy súlyozott kNN osztályozót használ, mely a szöveges információt és az annotátor korábbi annotációit egyaránt felhasználja az annotálási javaslatok kialakításához (Bikakis és mtsai, 2010). Az eddigiektől eltérően a Widlöcher és munkatársai (Widlöcher és Mathet, 2012) által fejlesztett Glozz eszköz nem automatikus annotálás segítségével, hanem a meglévő annotációk folyamatos monitorozási lehetőségével támogatja a konzisztens, jó minőségű annotációk készítését. Ehhez a fejlesztők egy GlozzQL-re keresztelt lekérdező nyelvet is készítettek.

A magyar nyelvű számítógépes nyelvészeti szakma, követve a nemzetközi gyakorlatot elsősorban annotátorok közötti egyezésmérést alkalmaz az annotáció minőségének ellenőrzésére. Ugyan a magyar szakirodalomban is találhatunk példát annotációk minőségének javítását célzó tanulmányokra (Novák, 2016), vagy már meglévő annotációk automatikus javítására (Kalivoda, 2017), a fentiekben bemutatott javaslattevő és annotációkat monitorozó alkalmazások használata tudomásunk szerint nem bevett gyakorlat.

5. Összegzés

Munkánk során bemutattuk, hogy az annotációk minősége jelentős mértékben befolyásolja a gépi tanuló algoritmusok teljesítményét, valamint javaslatot tettünk egy általunk fejlesztett annotációk konzisztenciájának fenntartását szolgáló eszköz alkalmazására. Kísérleteinkben egyetlen radiológus javítás előtti és utáni annotációja között mértünk intra-annotátor egyezést. A magas Cohen kappa és F1-mérték értékek arra utaltak, két annotáció jó egyezést mutat, azonban a modellünket a javítás előtt és utáni adatokon tanítva szembetűnő különbségeket tapasztaltunk. Az annotáció konzisztensebbé tételével 2-3%-os F1-mértékben tapasztalható javulást sikerült elérnünk az egyes névelemek esetén. Kísérleteink jó alapot szolgáltatnak egy későbbi, összetettebb rendszer fejlesztéséhez.

Köszönetnyilvánítás

Jelen kutatás az Innovációs és Technológiai Minisztérium ÚNKP-19-3 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült. A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM).

Hivatkozások

Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34, 555–596 (12 2008)

Bikakis, N., Giannopoulos, G., Dalamagas, T., Sellis, T.: Integrating keywords and semantics on document annotation and search. pp. 921–938 (01 2010)

Cai, T., Giannopoulos, A.A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K.K., Rybicki, F.J., Mitsouras, D.: Natural Language Processing Technologies in Radiology Research and Clinical Applications. RadioGraphics 36(1), 176–191 (jan 2016)

Cheng, L.T.E., Zheng, J., Savova, G.K., Erickson, B.J.: Discerning Tumor Status from Unstructured MRI Reports-Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing. Journal of Digital Imaging 23(2), 119–132 (apr 2010)

Dang, P.A., Kalra, M.K., Blake, M.A., Schultz, T.J., Stout, M., Lemay, P.R., Freshman, D.J., Halpern, E.F., Dreyer, K.J.: Natural Language Processing Using Online Analytic Processing for Assessing Recommendations in Radiology Reports. Journal of the American College of Radiology 5(3), 197–204 (mar 2008)

Dutta, S., Long, W.J., Brown, D.F., Reisner, A.T.: Automated Detection Using Natural Language Processing of Radiologists Recommendations for Additional Imaging of Incidental Findings. Annals of Emergency Medicine 62(2), 162–169 (aug 2013)

Farkas, R., Szarvas, Gy.: Eljárás radiológiai leletek automatikus BNO kódolására. In: V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007). p. 149–157. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2007)

Finlayson, M.A.: The story workbench: An extensible semi-automatic text annotation tool. In: Proceedings of the 4th Workshop on Intelligent Narrative Technologies. pp. 21–24 (2011)

Ford, E., Carroll, J.A., Smith, H.E., Scott, D., Cassell, J.A.: Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review. Journal of the American Medical Informatics Association 23(5), 1007–1015 (2016)

Ganchev, K., Pereira, F., Mandel, M., Carroll, S., White, P.: Semi-automated named entity annotation pp. 53–56 (06 2007)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997)

Ip, I.K., Mortele, K.J., Prevedello, L.M., Khorasani, R.: Focal Cystic Pancreatic Lesions: Assessing Variation in Radiologists' Management Recommendations. Radiology 259(1), 136–41 (apr 2011)

Kalivoda, Á.: Az igekötök gépi annotálásának problémái. In: Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből. pp. 100–108 (2017)

Kicsi, A., Pusztai, P., Szabó Ledenyi, K., Szabó, E., Berend, G., Vincze, V., Vidács, L.: Információkinyerés magyar nyelvű gerinc mr leletekből. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). p. 177–186. Szeged (2019)

Klebanov, B., Beigman, E.: From annotator agreement to noise models. Computational Linguistics 35, 495–503 (12 2009)

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. Yearbook of Medical Informatics pp. 44–128 (2008)

Morton, T., LaCivita, J.: Wordfreak: An open tool for linguistic annotation. (01 2003)

Muhie, S., Biemann, C., Eckart de Castilho, R., Gurevych, I.: Automatic annotation suggestions and custom annotation layers in webanno. pp. 91–96 (01 2014)

Novák, A.: Improving corpus annotation quality using word embedding models. Polibits 53, 49–53 (2016)

Oliveira, L., Gebeluca, C., Silva, A., Moro, C., Hasan, S., Farri, D.: A statistics and umls-based tool for assisted semantic annotation of brazilian clinical documents. pp. 1072–1078 (11 2017)

Pham, A.D., Névéol, A., Lavergne, T., Yasunaga, D., Clément, O., Meyer, G., Morello, R., Burgun, A.: Natural Language Processing of Radiology Reports for the Detection of Thromboembolic Diseases and Clinically Relevant Incidental Findings. BMC Bioinformatics 15(1), 266 (aug 2014)

Pons, E., Braun, L.M., Hunink, M.G., Kors, J.A.: Natural Language Processing in Radiology: A Systematic Review. Radiology 279(2), 329–343 (may 2016)

Raja, A.S., Ip, I.K., Prevedello, L.M., Sodickson, A.D., Farkas, C., Zane, R.D., Hanson, R., Goldhaber, S.Z., Gill, R.R., Khorasani, R.: Effect of Computerized Clinical Decision Support on the Use and Yield of CT Pulmonary Angiography in the Emergency Department. Radiology 262(2), 468–474 (feb 2012)

Rink, B., Roberts, K., Harabagiu, S., Scheuermann, R.H., Toomay, S., Browning, T., Bosler, T., Peshock, R.: Extracting Actionable Findings of Appendicitis from Radiology Reports Using Natural Language Processing. AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science p. 221 (2013)

Sistrom, C.L., Dreyer, K.J., Dang, P.P., Weilburg, J.B., Boland, G.W., Rosenthal, D.I., Thrall, J.H.: Recommendations for Additional Imaging in Radiology Reports: Multifactorial Analysis of 5.9 Million Examinations. Radiology 253(2), 453–61 (nov 2009)

Solti, I., Cooke, C.R., Xia, F., Wurfel, M.M.: Automated Classification of Radiology Reports for Acute Lung Injury: Comparison of Keyword and Machine Learning Based Natural Language Processing Approaches. In: Proceedings - 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2009. vol. 2009, pp. 314–319. NIH Public Access (nov 2009)

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: A Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics, Avignon, France (April 2012)

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical Information Extraction Applications: A Literature Review (jan 2018)

Widlöcher, A., Mathet, Y.: The glozz platform: a corpus annotation and mining tool. DocEng 2012 - Proceedings of the 2012 ACM Symposium on Document Engineering (09 2012)

Yetisgen-Yildiz, M., Gunn, M.L., Xia, F., Payne, T.H.: Automatic Identification of Critical Follow-Up Recommendation Sentences in Radiology Reports. AMIA Symposium pp. 1593–602 (2011)

Yim, W.w., Yetisgen, M., Harris, W.P., Kwan, S.W.: Natural Language Processing in Oncology. JAMA Oncology 2(6), 797 (jun 2016)

A tagmondati távolságszámítás módjainak hatása a névmási anaforafeloldásra

Kovács Viktória

Szegedi Tudományegyetem, Bölcsészettudományi Kar
viktoria.kovacs12@gmail.com

Kivonat: A névmási anaforafeloldás során a cél az egy szövegben található összes visszautaló névmás és az egyes visszautalásokhoz tartozó legközelebbi antecedensek azonosítása. A cikkben bemutatott gépi tanulási kísérletek segítségével a névmási visszautaló szó és az antecedense közötti távolság meghatározási módszereinek sikerességét vizsgálom. A vizsgálathoz a Szeged Koreferencia Korpusz (Vincze és mtsai, 2015) névmási visszautalásaiból építettem tanító és tesztfájlokat. Az osztályozók építéséhez a Szeged Korpuszban (Csendes és mtsai, 2005) található morfológiai és szintaktikai információkat használtam, a távolság meghatározásához állandó tényezőként pedig a Hobbs távolságot (Hobbs, 1978). A baseline modell ezeken kívül a két kifejezés közötti tagmondatzáró határátlépések számát vette figyelembe. A kísérletek során a tagmondatzáró határátlépések száma helyett olyan értékeket rendeltem a kifejezésekhez, amelyek kiszámításához figyelembe vettem a közbeékelődéseket (Gibson, 2000), az elérhetőségi elméletben megfogalmazott tagmondatok közötti kapcsolatokra vonatkozó elveket (Ariel, 2001), valamint azt a tényt, hogy a névmási visszautalások nagyrésze közelre történik (Grosz és mtsai, 1995).

1 Bevezetés

A koreferenciafeloldás egyik részfeladata a névmási anaforafeloldás, melynek során a cél a szöveg összes visszautaló névmásának és az egyes visszautalásokhoz tartozó legközelebbi antecedensek azonosítása. A kommunikáció során ezeknek a viszonyoknak a pontos felismeréséhez hozzájárulnak a morfológiai, a szintaktikai, a szemantikai és a pragmatikai információk is. A névmási anaforafeloldással kapcsolatos egyik leggyakoribb probléma a referenciális többértelműségen alapul, azaz a kifejezés több antecedensre is visszautalhat a szövegben: a jelen tanulmányban ismertetett vizsgálat a névmáshoz legközelebb eső antecedens azonosítását (Closest First) célozta meg. Egy másik gyakori probléma azon névmások kiszűrése, amelyek nem utalnak vissza. A magyar nyelvvel kapcsolatban (Lejtovicz és Kardkovács, 2006; Varasdi és mtsai, 2007; Miháltz, 2012) munkáit olvashatjuk, amelyek egyre jobb eredményeket érnek el a probléma megoldásának kapcsán.

Az anaforafeloldás során alapvető kérdés az antecedens keresési hatókörének meghatározása. Ezzel kapcsolatban két megközelítés lehetséges: vagy egy előre rögzített keresési tartományon belül vizsgáljuk meg a potenciális antecedensjelölteket és hasonlítjuk össze őket bizonyos grammatikai tulajdonságaik alapján; vagy

dinamikusan bővítjük a keresés hatókörét, és sorra vizsgáljuk a potenciális antecedensjelölteket addig, amíg nem találjuk meg a legvalószínűbb antecedensjelöltet. A két kifejezés közötti távolság tehát egy sarkalatos pontja mind a szabály alapú, mind a gépi tanuláson alapuló automatikus névmási anaforafeloldásnak. Ezen jellemző meghatározásához számos nyelvészeti elmélet is figyelembe vehető.

Jelen kísérlet egy korábbi munkán (Kovács, 2019) alapul, amelyben az elérhetőségi elmélet hatásának vizsgálata történt meg. A mostani kísérletben azt vizsgálom, hogy a szövegben is megjelenő névmáshoz tartozó antecedens azonosítása során a rögzített hatókör kiszámításának módja hogyan befolyásolja a gépi tanulás sikerességét. Ehhez különböző szintaktikai és kognitív nyelvészeti alapú elméletek távolságot befolyásoló tényezőit veszem figyelembe. Nem célom a jelen munkával az automatikus névmási anaforafeloldás problémáira megoldást találni, pusztán egyetlen jellemző meghatározási módjának hatását bemutatni egy gépi tanulási kísérleten keresztül. Az eredmények összefüggéseinek ismertetése, illetve nyelvészeti keretben való értelmezése további vizsgálatokat igényelnek.

1.1 Az anaforafeloldás kognitív nyelvészeti alapjai

Számos nyelvészeti modell, amely a koreferencia-, és ezen belül is az anaforafeloldást tűzte ki céljául, azon az elven alapul, hogy a beszélő vagy szövegalkotó a referáló kifejezés megválasztása során figyelembe veszi a hallgató vagy címzett mentális állapotát, hiszen az a célja, hogy a címzett felismerje azt, hogy mire utal, és ezáltal megértse a közölni kívánt információt (Ariel, 2001; Grósz és mtsai, 1995). Ez alapján a referáló kifejezés formájából és grammatikai tulajdonságaiból utólag is kikövetkeztethető a szövegben, hogy a szövegalkotó szerint mely entitások voltak az adott szituációban a címzett mentális állapotának középpontjában. A címzett mentális állapotának modellezése során azonban nem csak a kifejezések formája és grammatikai tulajdonságai lehetnek iránymutatók, hanem az adott kifejezés szövegben való elhelyezkedése is, hiszen minél nagyobb a távolság a visszautaló szó és az antecedense között, annál nagyobb erőfeszítésre van szüksége a címzettnek a kapcsolat feldolgozásához. A következő fejezetben azokat az elméleti megállapításokat fogom részletezni, amelyeket a két kifejezés közötti távolságra vonatkozóan a kísérletek során felhasználtam.

1.2 A visszautaló névmás és az antecedense közötti távolság

A két kifejezés közötti távolságból következtethetünk arra az erőfeszítésre, amelyet a címzettnek ki kell fejtenie ahhoz, hogy azonosítsa az anaforához tartozó antecedenst. Ennek megadásához általában két érték vehető figyelembe.

A Hobbs-távolság (Hobbs 1978) a két kifejezés közötti főnévi csoportok számát mutatja, azaz a lehetséges antecedensjelöltek számát, amelyekről a hallgatónak vagy címzettnek meg kell állapítania, hogy nem az anafora antecedensei. Ezt a jellemzőt minden egyes kísérletben figyelembe vettem kiindulási alapként.

A másik érték a két kifejezés közötti tagmondatok számából származik, ennek a jellemzőnek a hatását vizsgálja a három kísérlet. A tagmondatok számára hagyományos módon úgy tekintünk egyszerűen, mint egy hatókörre, ezt az elvet követi a baseline modellként szolgáló első kísérlet is. Azonban kognitív nyelvészeti kutatások alapján megállapítható, hogy nem pusztán a névmás és az antecedense közötti tagmondatok száma, de a tagmondatok egymáshoz való viszonya is meghatározza az anaforafeloldáshoz szükséges erőfeszítés mértékét.

A kognitív nyelvfeldolgozás során a feladat a szavak értelmezése és szerkezetbe való beépítése. Ezt a feldolgozást azonban nehezíti, amikor egy szerkezetbe egy újabb szerkezet közbeékelődik, és a korábbi szerkezetet a címzettnek hiányos állapotában tárolnia kell mindaddig, amíg a közbeékelt szerkezet teljessé nem válik. Minél több ilyen közbeékelődés található egy mondatban, annál nehezebb a címzettnek feldolgoznia az információt. Ebből az a következtetés vonható le, hogy az ilyen típusú mondatok értelmezése során a címzettnek nagyobb erőfeszítésre van szüksége az értelmezéshez, mint az egyszerű tagmondatok értelmezése során, és ez a különbség hatással van az antecedens azonosításához szükséges erőfeszítésre is (Gibson, 2000).

Szintén a hallgató és a címzett mentális állapotát helyezi a középpontba az elérhetőségi elmélet (Ariel, 2001), amely kitér az anafora és az antecedense közötti távolság anaforafeloldásra gyakorolt hatására is. Az elmélet szerint az alárendelő tagmondatból kisebb erőfeszítéssel érhető el az anaforához tartozó antecedens, mint a mellérendelő tagmondatból. Ezt a két jelenséget figyelembe véve végeztem el a második kísérletet.

A harmadik megállapítás, amelyet figyelembe vettem a kísérlet során, azt mondja ki, hogy a névmással való visszautalás azt feltételezi, hogy az antecedens könnyen elérhető, ezért a névmás antecedense a névmással azonos vagy szomszédos mondatban helyezkedik el (Grosz és mtsai, 1995). Azokat a névmásokat, amelyek ennél messzebbre utalnak vissza *long distance*, azaz nagy hatókörű anaforának nevezi a szakirodalom, és jellemzően olyan kifejezésre utalnak vissza, amely diskurzustopik, tehát a diskurzus fő témája, illetve annyira szaliens entitás, mint a szövegalkotó vagy a címzett maga. Ezzel az elvvel kiegészítve végeztem el a harmadik kísérletet.

2 Korpusz

A kísérletet a Szeged Korpusz (Csendes és mtsai, 2005) koreferencia-annotált alkorpuszán, a Szeged Koreferencia Korpuszon (Vincze és mtsai, 2015) végeztem el, ami iskolai fogalmazásokból és újsághírekből áll. Az összes visszautalás közül kizárólag a szövegben is megjelenő névmási visszautalásokat gyűjtöttem ki (tehát nem vettem figyelembe a zérókat), ez összesen 725 visszautalást jelent.

2.1 Módszer

A gépi tanuláshoz a Mention-Pair (Soon és mtsai, 2001) modellt használtam, amely során lehetséges visszautaló névmások és a hozzájuk tartozó lehetséges antecedensjelöltekből álló párokat nyertem ki a korpuszból. Ezek a párok és a hozzájuk rendelt morfológiai és szintaktikai jellemzők adták a tanító és tesztfájlokat.

A modell előnye, hogy a párokhoz kézzel hozzáadott jellemzők tanulásra gyakorolt hatása egymástól függetlenül vizsgálható, tehát a különböző elméletekben megfogalmazott elvek automatikus anaforafeloldásra gyakorolt hatásai ellenőrizhetők. A korpuszban nem csak a koreferens kapcsolatok találhatók meg, hanem a kifejezésekhez tartozó morfológiai és szintaktikai elemzések is. A konstituenselemzés segítségével a főnévi csoportok, valamint a hozzájuk tartozó morfológiai elemzések kinyerhetők a fájlokból. A párok első eleme olyan főnévi csoport, amely a korpuszban a morfológiai elemzés során PRON címkét kapta. Az antecedensjelöltek a névmásokat a szövegben megelőző főnévi csoportok (NP).

A tanító fájlok úgy jöttek létre, hogy minden egyes lehetséges visszautaló névmáshoz párként hozzárendeltem az őt megelőző NP-ket a kézzel is annotált valódi antecedensével bezárólag. Tehát minden esetben annyi pár jött létre, ahány NP található a névmás és az antecedense között, ezek a negatív példák, plusz egy, maga az antecedense, ez a pozitív példa. Mivel a távolság hatását szerettem volna vizsgálni, ezért minden annotált névmáshoz csak egy pozitív példa lett hozzárendelve, a szövegben hozzá legközelebbi. Azon névmások esetében, amelyekhez nem volt antecedens annotálva, tehát amelyek nem utaltak vissza, a névmást megelőző három főnévi csoporttal alkottam párokat. Ezzel azokhoz a névmásokhoz is generáltam negatív példákat, amelyek nem utaltak vissza a szövegben.

A tesztfájlokban viszont minden névmáshoz hozzárendeltem párként minden őt megelőző főnévi csoportot a szöveg első főnévi csoportjával bezárólag, hiszen elvben bármely főnévi csoport antecedense lehet bármely névmásnak. Így a tesztfájlokban a pozitív példák átlagosan az összes pár 0,39%-át tesztik ki.

A tanító fájlban való negatív példák szűrésére azért volt szükség, mert egy szövegben arányosan sokkal több a negatív példa, mint a pozitív, ez alapján pedig a legtöbb osztályozó a tesztfájban az összes párt negatívnak ítéli. A szűrés után a tanító fájlokban a pozitív példák száma az összes pár 10,35%-a lett.

2.2 A tanuláshoz használt jellemzők

A névmásokból és antecedensjelöltekből álló párokhoz hozzárendeltem a két kifejezésre vonatkozó morfológiai és szintaktikai információkat. A kísérlet célja az volt, hogy a párok tagjai közötti tagmondati távolság (*CPdist* jellemző) tanulásra gyakorolt hatását megvizsgáljam. Jelenleg a tanító fájlokban 14 tényező jellemzi a névmási anafora és antecedens párokat, ezek mind a Szeged Korpuszból származnak. A tanító fájlba maguk a kifejezések nem kerülnek bele, kizárólag az őket jellemző tulajdonságok.

A tanító fájlban a párok a következő módon vannak jellemezve (1):

(1) [CPdist, NPdist, antPOS, anaTyp, anaCas, antCas,
casAgr, anaNum, antNum, numAgr, anaPer, antPer, perAgr, anaphoric]

A jellemzők két fő csoportra oszthatók. Az első csoportba morfológiai és szintaktikai jellemzők kerültek, amelyeket közvetlenül a korpuszban a szavakhoz rendelt címkékből nyertem ki. A másik nagyobb csoportba azok a jellemzők tartoznak, amelyeket szintén a korpuszból, de nem a hozzárendelések segítségével

nyertem ki. Az utolsó jellemző pedig a koreferencia annotációból származó adat, amely azt mutatja, hogy a pár anaforikus-e vagy sem. A következő két fejezetben ezen jellemzők részletes ismertetése található, bővebben olvasható róluk Kovács 2019 munkájában (Kovács 2019).

2.2.1 A tanuláshoz felhasznált morfológiai és szintaktikai tulajdonságok

Az *antPOS* jellemző a magyarlánc elemzésében az antecedensjelölt fejéhez rendelt POS Taget írja ki értékként.

Az *anaTyp* jellemző a névmás típusát jelöli. A magyarlánc morfológiai elemzőjének PronType típusú címkéit veheti fel értékként.

Az antecedensjelölt esete *antCas*, száma *antNum*, és személye *antPer* három különböző jellemzőként jelenik meg a tanító fájlban, és szintén a morfológiai elemzésből származik. A főnévi csoportnál a csoport fejéhez rendelt Case típusú morfológiai címkével egyezik meg az érték.

Az anaforikus névmás esete *anaCas*, száma *anaNum*, személye *anaPer* szintén három különböző jellemző, amelyek a morfológiai elemzésből származnak.

Az egyeztetés eset szerint *casAgr*, szám szerint *numAgr*, személy szerint *perAgr* jellemzők azt vizsgálják meg, hogy a névmáshoz rendelt eset, szám és személy címke és az antecedenshez rendelt eset, szám és személy címke megegyezik-e. Abban az esetben, ha megegyezik a jellemző, az 1-es értéket veszi fel, ha pedig nem, a 0-t.

2.2.2 A tanuláshoz felhasznált távolságra vonatkozó tulajdonságok

A jellemzők meghatározása során azzal kísérleteztem, hogy a tagmondati távolságot a korábban említett elveket is figyelembe véve többféleképpen határoztam meg. Végül három esetet különböztettem meg. Az első esetben baseline-ként a *CPdist* jellemző egyszerűen a két kifejezés közötti tagmondatzáró határátlépéseket jelentette. Azokat az eseteket, ahol több záró határátlépés egybeesett, egy határátlépésként tekintettem. A baseline *CPdist* értéke az 2. és 3. példában szereplő visszautalás esetében így 5 lett.

2. [[Amíg vártuk **Petit**, [mert úgy hívják a kocsis haveromat], elmentünk fagyizni], [ott meg találkoztunk a barátom haverjaival.]] [Ők is ép fagyiztak.][[Velük elbeszélgettünk], [aztán jött **ő**]és [mentünk Tófaluba]].

3. [[Amíg vártuk **Petit**, [mert úgy hívják a kocsis haveromat], elmentünk fagyizni], [ott meg találkoztunk a barátom haverjaival.]] [Ők is ép fagyiztak.][[Velük elbeszélgettünk, [aztán jött **ő** is .]]

A második esetben figyelembe vettem közbeékelődéseket és az elérhetőségi elméletben az alá- és mellérendelő tagmondatok kapcsán megfogalmazott alapelveket. A szakirodalom alapján közbeékelődéseknek azokat az eseteket tekintettem, ahol a közbeékelt mondatnak sem a kezdete, sem a vége nem esik egybe az őt tartalmazó mondat kezdetével vagy végével. A 2. példában a *mert úgy hívják a kocsis haveromat* tagmondat megszakítja az *Amíg vártuk Petit (...), elmentünk fagyizni*. teljes tagmondatot. Tehát azt az egységet, hogy *Amíg vártuk Petit,* ebben a formában, hiányosan kell tárolnia a hallgatónak, mindaddig, amíg a közbeékelt mondat végéhez nem ér. Ezért a visszautaló névmástól számítva a tagmondati határátlépések számát

úgy vettem figyelembe, hogy a közbeékelődött mondat esetében egy belépési és egy kilépési értéket is számításba vettem.

Alárendelésnek tekintettem azokat az eseteket, ahol a beágyazott mondat kezdete vagy vége egybeesett az őt tartalmazó mondat kezdetével vagy végével, ezekben az esetekben a határátlépés egy pontot ért. Mellérendelésnek pedig azokat a szerkezeteket tekintettem, amelyeket más mondat tartalmazott, és ahol a megelőző mondat vége és a soron következő mondat eleje közé nem került főnévi csoport: itt a határátlépés két ponttal növelte a *CPdist* jellemző értékét. Ezeknél a szerkezeteknél is egy határátlépésnek számítottak az egybeeső mondatkezdő vagy egybeeső mondatzáró határok. A 2. példa értéke így 8 lett, a 3. példáé pedig 7.

A harmadik eset annyiban tért el a másodiktól, hogy a teljes mondat határátlépések, tehát azok a mondatok, amelyeket nem tartalmaz más mondat, nem egy, hanem három pontot értek, ezzel a nagy hatókörű anaforák esetét igyekeztem pontosítani. Ez esetben a 2. példa a 12 értéket vette fel, a 3. példa pedig 11-et.

3 A gépi tanulási kísérletek

A meghatározott jellemzők alapján a tanítófájlokon a Random Forest (Breiman, 2001) algoritmussal építettem osztályozót a Weka szoftver (Eibe és mtsai, 2016) segítségével. Mivel a cél egy adott jellemző meghatározási módszerének tanulásra gyakorolt hatásának vizsgálata volt, ezért az osztályozó építése során a Wekának az algoritmussal kapcsolatos alapértelmezett beállításain nem változtattam. A névmási anaforafeloldással kapcsolatban az osztályozóról olvashatók további eredmények a baszk (Arregi és mtsai, 2010) a maláj (Xian és mtsai, 2016) és az orosz (Ionov és Kutuzov, 2014) nyelveken végzett kísérletekről is.

3.1 Az osztályozó tesztelése

A három osztályozó teszteléséhez az alacsony számú visszautalás miatt a keresztvalidálás módszerét alkalmaztam. A korpuszt a szövegek alapján tíz részre osztottam: kilenc részből készült el a tanító fájl, egy részből pedig a tesztfájl. Ezt a módszert pedig tízszer megismételtem, a végleges kiértékeléshez pedig az egyes tesztek átlagát használtam fel. A tíz tesztfájlban található névmási visszautalások arányait a 1. táblázat mutatja.

A tesztfájlokon a kiértékelés során a fals pozitív példák szűréséhez a closest-first módszert (Soon és mtsai. 2001) alkalmaztam, mivel a távolság hatását vizsgáltam. Tehát a névmáshoz a szövegben legközelebb álló, az osztályozó által pozitívnak ítélt antecedensjelöltet tekintettem egyedül a névmáshoz rendelt antecedensnek. Ezzel egyre csökkentettem minden névmás tekintetében a pozitív példák számát, a legközelebbire.

1. táblázat: A tesztfájlok adatai (Dem= mutató névmás, Prs= személyes névmás, Rel= vonatkozó névmás, Other= egyéb névmási visszautalás)

	Dem	**Prs**	**Rel**	**Other**	**Összesen**
TEST1	7	22	44	1	74
TEST2	20	12	38	0	70
TEST3	12	22	42	0	76
TEST4	11	24	37	1	73
TEST5	12	25	41	0	78
TEST6	5	22	36	0	63
TEST7	10	17	39	1	67
TEST8	10	25	32	0	67
TEST9	12	22	40	0	74
TEST10	13	18	50	2	83
Összesen	112	209	399	5	725

2. táblázat: A tanulási kísérletek adatai (P = precision, R = recall, F = F-measure)

	Baseline			**Exp1**			**Exp2**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	22,41	35,14	27,37	22,31	36,49	27,69	23,53	37,84	29,02
TEST2	28,07	45,71	34,78	29,66	50,00	37,23	32,14	51,43	39,56
TEST3	29,20	43,42	34,92	28,57	42,11	34,04	30,63	44,74	36,36
TEST4	37,50	45,21	40,99	34,83	42,47	38,27	38,46	47,95	42,68
TEST5	40,19	55,13	46,49	39,62	53,85	45,65	41,18	53,85	46,67
TEST6	31,65	39,68	35,21	35,62	41,27	38,24	35,82	38,10	36,92
TEST7	36,61	61,19	45,81	41,84	61,19	49,70	39,60	59,70	47,62
TEST8	39,02	47,76	42,95	38,55	47,76	42,67	40,74	49,25	44,59
TEST9	30,85	39,19	34,52	34,04	43,24	38,10	34,02	44,59	38,6
TEST10	41,75	51,81	46,24	37,72	51,81	43,65	51,81	51,81	51,81
ÁTLAG	**33,73**	**46,42**	**38,93**	**34,28**	**47,02**	**39,52**	**36,79**	**47,92**	**41,38**

4 Eredmények

Ahhoz, hogy megtudjam, hogy az általam alkalmazott tagmondati távolságszámítás eredményes-e a gépi tanulás során, három tesztet végeztem el. A Baseline tesztelése

során nem tettem különbséget a tagmondatok között, ezek az eredmények láthatók a 2. táblázat Baseline oszlopában. Az első tesztelésnél már figyelembe vettem a közbeékelődéseket és az alá- és mellérendelő mondatok közötti különbségeket, ezt mutatja a táblázatban az Exp1 oszlop. A második teszt során már a nagy hatókörű anaforák alapján megfogalmazott elveket is figyelembe vettem, ezt mutatja az Exp2 oszlop.

Mivel az Exp2 eredményei konzisztensen jobbak lettek a Baseline eredményeinél, azt is megvizsgáltam, hogy az egyes visszautalási típusok tekintetében mekkora a változás. Mivel a visszaható névmási visszautalások száma a korpuszban kevés volt, ezért a mutató névmási (Dem), személyes névmási (Prs) illetve vonatkozó névmási (Rel) kategóriákat vizsgáltam meg. Ezeknek az eredményeit a 3. és a 4. táblázat mutatja.

3. táblázat: A tanulási kísérletek adatai a visszautaló névmások típusai szerint a Baseline esetében (P = precision, R = recall, F = F-measure)

	Dem			**Prs**			**Rel**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	01,96	14,29	03,45	50,00	04,54	08,33	39,34	54,55	45,71
TEST2	06,82	15,00	09,38	0	0	0	44,62	76,32	56,31
TEST3	04,08	16,67	06,56	33,33	13,64	19,35	51,85	66,67	58,33
TEST4	0	0	0	40,00	25,00	30,77	63,41	70,27	66,67
TEST5	11,90	41,67	18,52	20,00	04,00	06,66	61,67	90,24	73,27
TEST6	0	0	0	20,00	04,54	07,40	60,00	66,67	63,16
TEST7	06,67	30,00	10,91	66,67	23,53	34,78	57,63	87,18	69,39
TEST8	11,54	30,00	16,67	40,00	16,00	22,86	55,56	78,13	64,94
TEST9	03,70	08,33	05,13	33,33	04,54	08,00	42,86	67,50	52,43
TEST10	04,17	07,69	05,41	33,33	05,55	09,52	54,79	80,00	65,04
ÁTLAG	**5,08**	**16,36**	**7,60**	**33,67**	**10,14**	**14,77**	**53,17**	**73,75**	**61,52**

5 Kiértékelés, hibaelemzés

Az összes visszautalást tekintve az Exp1 kísérletben a tíz teszt átlaga alapján javított az eredményeken az új *CPdist* számolási módszer a Baseline-hoz képest, a pontosságon 0,55%-kal, a fedésen 0,59%-kal, az F mértékek átlagait tekintve pedig 0,6%-kal. Ennek ellenére nem vonható le egyértelműen az a konklúzió, hogy az új jellemző eredményesebb, mivel 5 teszt javított, 5 pedig rontott a Baseline *CPdist* számolási módszeréhez képest. Az Exp2 kísérletben minden egyes tesztről elmondható, hogy jobb eredménye lett, mint a Baseline tesztjeinek. Az összes teszt átlaga alapján a fals pozitív esetek szűrésén javított az új jellemző a legtöbbet, tehát a pontosságon 3,07%-kal. A fedésen 1,5%-kal, az F mértéken pedig 2,45%-kal javított az új jellemző, tehát egyértelműen ez a legeredményesebb a három számítási módszer közül.

4. táblázat: A tanulási kísérletek adatai a visszautaló névmások típusai szerint az Exp2 esetében (P = precision, R = recall, F = F-measure)

	Dem			**Prs**			**Rel**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
TEST1	03,85	28,57	06,78	20,00	04,54	07,41	41,67	56,82	48,08
TEST2	13,95	30,00	19,05	25,00	08,33	12,50	46,77	76,32	58,00
TEST3	04,08	16,67	06,56	42,86	13,64	20,69	53,70	69,05	60,42
TEST4	03,03	09,09	04,55	53,85	29,17	37,84	63,41	70,27	66,67
TEST5	12,20	41,67	18,87	25,00	04,00	06,89	63,16	87,80	73,47
TEST6	0	0	0	33,33	04,54	08,00	63,89	63,89	63,89
TEST7	07,89	30,00	12,50	66,67	23,53	34,78	61,11	84,62	70,97
TEST8	13,04	30,00	18,18	45,45	20,00	27,78	54,35	78,13	64,10
TEST9	04,00	08,33	05,41	33,33	09,09	14,29	46,15	75,00	57,14
TEST10	02,94	07,69	04,26	50,00	05,56	10,00	56,34	80,00	66,12
ÁTLAG	**6,50**	**20,20**	**9,61**	**39,55**	**12,24**	**18,02**	**55,06**	**74,19**	**62,88**

Az egyes visszautalási típusok szerint is elmondható, hogy az Exp2 *CPdist* jellemzője átlagosan javított az eredményeken. A legnagyobb javulás a személyes névmási visszautalásnál látható, ahol a fedésen 2,11%-kal, a pontosságon 5,88%-kal, az F mértéken pedig 3,25%-kal javított a jellemző. A mutató névmási visszautalás esetében a fedésen 3,84%-ot, a pontosságon 1,41%-ot az F mértéken 2,01%-ot javított. A legkisebb változás a vonatkozó névmási visszautalásnál történt, ebben az esetben a fedés 0,44%-kal, a pontosság 1,88%-kal, az F mérték pedig 1,36%-kal nőtt. Ennek oka, hogy a vonatkozó névmási visszautalás antecedense mindig a megelőző tagmondatban található, ezért az új jellemző nem változtatott számottevően az értékeken. Azonban azokban az esetekben, ahol a visszautalás messzebbre is történhet, a távolság számítás módszerének finomítása javított az eredményeken, a személyes névmási visszautalás esetében a fals pozitív, a mutató névmási visszautalás esetében pedig a fals negatív előfordulások szűrésében.

Általánosságban elmondható, hogy a tesztek között nagy eltérések vannak, ennek oka, hogy az egyes tesztekben eltér a visszautalások száma. Emellett az is szembetűnő, hogy a vonatkozó névmási visszautalás felismerése sokkal eredményesebb, mint a személyes- vagy mutató névmási visszautalásé. Ennek egyik oka, hogy a névmás és antecedense közötti nagyobb távolság esetén az osztályozó gyakran tévesen egy közelebbi főnévi csoportot jelöl meg antecedensnek. A másik ok pedig maguknak a visszautalásoknak az aránya a szövegekben. A vonatkozó névmási visszautalások sokkal gyakoribbak a szövegekben, ezért mind a tanító, mind a tesztfájlokban nagyobb arányban fordulnak elő.

Általános problémát jelent azoknak a névmásoknak a kezelése, amelyeknek nem szükséges a szövegben antecedenst keresni. Mivel jelen tanulmány célja kizárólag a

távolságszámítás hatásának vizsgálata volt, ezt a problémát nem kezeltem. Jelenleg az osztályozó az összes PRON címkével jellemzett kifejezést kigyűjti, és sorra vizsgálja hozzájuk a lehetséges antecedensjelölteket. Az ebből a problémából fakadó hibákat úgy lehetne orvosolni, hogy a tanító fájlhoz olyan negatív példákat adok, amelyek deiktikus névmásokat tartalmaznak, azonban ezzel tovább növelném a negatív és pozitív példák közti arányok különbségét. Ahhoz hogy a későbbiekben az egyes elméletekből fakadó jellemzők tanításra gyakorolt hatásáról pontosabb képet kapjak a tesztfájlokban szereplő névmások, illetve a negatív példák előszűrése lesz szükséges.

6 Konklúzió

A gépi tanulási kísérlet célja az volt, hogy megvizsgáljam a visszautaló névmás és az antecedense közötti távolság számítási módszereinek hatását egy osztályozó eredményességére. Az Exp2 kísérlet eredményei alapján elmondható, hogy a két kifejezés közötti mondatszintű távolság számítása során javít az osztályozó eredményességén az, ha a különböző mondattípusokat más-más súllyal vesszük számításba, leginkább azokban az esetekben, ahol az antecedens nem azonos vagy szomszédos tagmondatban található.

Hivatkozások

Ariel, M.: Accessibility theory. An overview. In Sanders, T., Schilperoord, J., Spooren W. (szerk.) Text Representation: Linguistic and Psycholinguistic Aspects. Amsterdam: John Benjamins Publishing Company. 29–87. (2001)

Arregi, O., Ceberio, K., Díaz de Illarraza, A., Goenaga, I., Sierra, B., Zelaia, A.: A First Machine Learning Approach to Pronominal Anaphora Resolution in Basque. In Kuri-Morales, A., Simari, G. R. (szerk.) Advances in Artificial Intelligence – IBERAMIA 2010. Vol. 6433 Springer Berlin Heidelberg. 234–243. (2010.)

Breiman, L.: Random Forest. Machine Learning 45/1:5–32. (2001)

Csendes D., Csirik J., Gyimóthy T., Kocsor A.: The Szeged Treebank. In: Matoušek, V., Mautner, P., Pavelka, T. (szerk) Text, Speech and Dialogue. TSD 2005. Lecture Notes in Computer Science, vol 3658. Springer, Berlin, Heidelberg (2005)

Eibe, F., Hall, M. A., Witten, I. H.: The WEKA Workbench. Online Appendix for „Data Mining: Practical Machine Learning Tools and Techniques". Fourth Edition. Morgan Kaufmann. (2016)

Gibson, E.: The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., O'Neil, W. (szerk.) Image, language, brain, Cambridge, MA: MIT Press, 95–126. (2000)

Grosz, B. J., Joshi, A. K., Weinstein, S.: Centering: A Framework for Modeling the Local Coherence of Discourse Computational Linguistics, Volume 21, Number 2, 203–225. (1995)

Hobbs, J. R.: Resolving pronoun references, Lingua, Volume 44, Issue 4, 311-338. (1978)

Ionov, M., Kutuzov, A. The impact of morphology processing quality on automated anaphora resolution for Russian. In Selegey, V.P., Baytin, A.V., Belikov, V.I., Boguslavsky, I.M., Dobrov, B.V., Dobrovolsky, D.O., Zakharov, L.M., Iomdin, L.L., Kobozeva, I.M., Kozerenko, E.B., Krongauz, M.A., Laufer, N.I., Lukashevich, N.V., McCarthy, D., Nivre,

J., Osipov, G.S., Raskin, V., Hovy, E., Sharoff, S.A. (szerk.) Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" 232-240. (2014)

Kovács, V.: Az elérhetőségi elmélet névmási anaforafeloldásra gyakorolt hatása. In Váradi, T. (sorozatszerkesztő), Ludányi, Zs., Gráczi, T. E. (szerk.) Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019. XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia. Budapest: MTA Nyelvtudományi Intézet. 114–123. (2019)

Lejtovicz, K. E., Kardkovács, Z. T.: Anaforafeloldás magyar nyelvű szövegekben. In Alexin, Z., Csendes, D. (szerk.) IV. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2006 362–363. Szegedi Tudományegyetem. Szeged (2006)

Miháltz, M.: Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. Általános Nyelvészeti Tanulmányok, 24, 151–166. (2012)

Soon, W. M., Ng, H. T., Lim, D. C. Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27:521–544. (2001)

Varasdi, K., Vajda, P., Miháltz, M., Naszódi, M.: NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével. In Tanács, A., Csendes, D. (szerk.) V. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2007. 138–146 Szegedi Tudományegyetem. Szeged (2007)

Vincze, V., Hegedűs, K., Farkas, R.: SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged, 312–319 (2015)

Xian, B. C. M., Saloot, M. A., Ghazali, A. S. M., Bouzekri, K., Mahmud, R. Lukose, D. Benchmarking Mi-AR: Malay anaphora resolution. In *2016 International Conference on Optoelectronics and Image Processing (ICOIP)*, IEEE, Warsaw 59-69 (2016)

KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése

Vadász Noémi

Nyelvtudományi Intézet
Budapest, Benczúr utca 33.
e-mail: vadasz.noemi@nytud.hu

Kivonat A cikk egy többrétegű, kézzel annotált korpuszt ismertet, bemutatja annak elemzési rétegeit – különös tekintettel az anafora- és koreferenciaannotációra – és az építés fázisait, valamint felvillantja a felhasználási lehetőségeket. A korpusz szabadon elérhető és felhasználható, az építéshez használt eszközök és dokumentációik, valamint az annotálási útmutatók biztosításával pedig lehetőség nyílik annak további szövegekkel történő bővítésére.
Kulcsszavak: korpusz, annotálás, anafora, koreferencia

1. Háttér

A KorKorpusz tervezésekor a jelenleg létező legnagyobb magyar koreferenciakorpusz nyújtott inspirációt[1]. A SzegedKoref (Vincze és mtsai, 2015) a Szeged Korpusz (Csendes és mtsai, 2005) egy részét felhasználva készült, újsághíreket és iskolai fogalmazásokat láttak el koreferenciaannotációval. A legutóbbi publikáció alapján a SzegedKoref (Vincze és mtsai, 2018) 400 szöveget, 4 021 mondatot és 55 763 tokent tartalmaz. A szövegekben 2 456 anaforikus láncot[2] jelöltek meg.

Mi szükség van a SzegedKoref mellett még egy magyar koreferenciakorpuszra? Ez a kérdés több irányból is megközelíthető. A kézzel annotált, jó minőségű adat nagyon értékes erőforrás, és minél több van belőle, annál jobb. A cikkben ismertetett KorKorpusz összes elemzési rétege – a SzegedKorefhez hasonlóan – kézzel ellenőrzött minőségű, így nem csak az anafora- és koreferenciaannotáció hasznosítható belőle, hanem a többi nyelvi elemzés is. A két korpusz elemzési rétegei között azonban vannak különbségek, míg a SzegedKoref az MSD morfológiai kódkészlet[3] (Erjavec, 2004) egy feature-value formában megfogalmazott verzióját[4] használja morfológiai címkekészletként, addig a KorKorpusz morfológiai

[1] A magyar nyelvű koreferenciakorpuszok között meg kell említeni a Miháltz és mtsai (2007) és Miháltz (2012) tudásalapú koreferenciafeloldó rendszerének kiértékeléséhez használt korpuszokat. Ezek általános iskolai történelemkönyvből vett szövegekből állnak, amelyekben kézzel annotálták a különböző típusú anaforikus- és koreferenciakapcsolatokat. A korpuszokon egy annotátor dolgozott és a fent hivatkozott cikkek részletesen leírják az annotált típusokat, ám a korpuszok sajnos nem hozzáférhetőek.

[2] Az anaforikus láncok magukban foglalják a névmási anaforikus kapcsolatokat és a koreferenciaviszonyokat is.

[3] http://nl.ijs.si/ME/Vault/V3/msd/msd.pdf

[4] https://github.com/dlt-rilmta/panmorph/blob/master/panmorph_conll.pdf

rétege emMorph[5] (Novák és mtsai, 2017) és UD-kompatibilis[6] morfológiai címkéket tartalmaz. Egy másik különbség, hogy a SzegedKoref összetevős elemzést, míg a KorKorpusz dependenciaelemzést használ szintaktikai elemzésre. Végül a KorKorpusszal szemben a SzegedKoref nem tartalmaz tövesítést.

A fent említett különbségek ellenére elképzelhető a két korpusz együttes használata, így a SzegedKoref kb. 55 ezer tokenje kiegészülhet a KorKorpusz mindenkori tartalmával. Ehhez csupán az eltérő formátumú koreferenciaannotációt kell egységes formára hozni. Ugyanakkor fontos megemlíteni, hogy a KorKorpusz tervezésekor bizonyos elméleti kérdésekben másképp döntöttünk, mint a SzegedKoref-ben (például a KorKorpuszban az infinitívusz alanya is megjelenik zéró névmásként, beillesztettük a zéró létigéket és az elliptált igéket, jelöljük az általános alanyokat is, különválasztottuk az anaforikus kapcsolatokat a koreferenciaviszonyoktól stb).

A korpusz tervezésekor szem előtt tartottunk a könnyű elérhetőséget, használhatóságot és továbbfejleszthetőséget. A SzegedKoref engedélykérés után kutatási és oktatási célokra felhasználható, míg a KorKorpusz az összes dokumentáció és útmutató társaságában CC-BY-4.0 licensszel elérhető, így bárki továbbfejlesztheti és publikálhatja az eredményeit.

2. Anafora és koreferencia

Az információkinyerés és a kivonatolás területein számos olyan feladat van, amelyek megoldásához anafora- vagy koreferenciafeloldásra van szükség. Egy ilyen kapcsolatokat is tartalmazó korpusz hasznos erőforrás, akár tanítóanyagként, akár a kiértékelés során. Ám a szöveget átszövő kapcsolatok annotálása is nagyobb kihívást jelent.

A koreferencia és az anafora fogalma gyakran összemosódik, hiszen mindkét kapcsolattípus feloldása feltétele a szöveg pontos interpretációjának. Szem előtt kell tartani ugyanakkor a köztük lévő fontos különbségeket. Amint van Deemter és Kibble (1999) rávilágít, nem mindig világos, hogy az egyes korpuszok esetében pontosan mit értenek koreferencia alatt. Felhívja a figyelmet arra is, hogy míg a koreferencia szimmetrikus és tranzitív kapcsolat, addig az anafora nem, viszont az anafora kontextusfüggő.

A kétféle kapcsolat közötti különbség a KorKorpusz anafora- és koreferenciaanotálásánál is megmutatkozott, a részleteket lásd a 4.7. és a 4.8 fejezetekben.

3. A korpusz főbb adatai

A KorKorpusz az `e-magyar` újabb verziójának (Indig és mtsai, 2019) keretében használt formátumot[7] követi, amelyben a tokenek soronként szerepelnek és a

[5] `https://e-magyar.hu/en/textmodules/emmorph_codelist`
[6] `https://github.com/dlt-rilmta/panmorph/blob/master/panmorph_ud.pdf`
[7] `https://github.com/dlt-rilmta/xtsv`

mondatokat üres sor választja el egymástól. A különböző elemzési rétegek tabbal elválasztott oszlopokban kapnak helyet, az oszlopok sorrendje nem kötött, azt a fájl első sorában található fejléc határozza meg. A korpusz jelenleg 95 dokumentumot, 1 436 mondatot és 31 492 tokent tartalmaz, amelybe beleszámítanak az írásjelek és a zéró elemek is[8].

Jelenleg két forrásból gyűjtött szövegeket tartalmaz a korpusz, amelyeket az OPUS gyűjteményéből (Tiedemann, 2012) válogattuk. A szövegek egy részét a magyar Wikipédiáról gyűjtöttük, másrészt a GlobalVoices hírportál[9] magyar nyelvre lefordított hírei közül válogattunk. A KorKorpusz örökli ezeknek a forrásoknak a nyílt hozzáférhetőségét.

A korpusz az építéséhez készített eszközökkel és dokumentációikkal, valamint az annotálási útmutatókkal együtt az alábbi GitHub repozitóriumban érhető el: `https://github.com/vadno/korkor_pilot`.

4. A korpuszépítés lépései

A korpusz tervezésekor a munkát egy feldolgozó láncolatként képzeltük el. Célunk volt, hogy minél több lépést automatizáljunk és emberi munkát csak az eszközök kimenetének javításához használjunk.

Bizonyos elemzési lépésekhez az `e-magyar` második verzióját (Indig és mtsai, 2019) használtuk, amelyre ezentúl a cikkben munkanevén, `emstv`-ként hivatkozunk. Mivel az `emtsv` egy szövegfeldolgozó pipeline, ahol egy adott elemzési lépés kimenete a következő lépés bemenetét képezi, így nem lenne hatékony csupán a – jelen bemutatott korpusz szempontjából – legutolsó lépés után kézzel javítani a kimenetet, hiszen addigra a korábbi lépésekben keletkezett hibák hógolyóként még több hibát görgetnének maguk előtt. Így többlépéses kézi javítást alkalmaztunk, amely ugyan idő- és munkaigényes feladat, viszont könnyebben kontrollálható. Az `emtsv` kimenetét két körben volt szükséges ellenőrizni és javítani, ezután a saját eszközeink (zérónévmás-beszúró, anaforafeloldó) kimenetét is ellenőrizni kellett.

Az alábbi felsorolás tartalmazza az egyes elemzési és ellenőrzési lépéseket. Zárójelben a használt eszközök neve jelenik meg (a saját fejlesztésű eszközök **vastag betűvel** kiemelve).

1. szöveggyűjtés
2. elemzés (emtsv/emToken, emtsv/emMorph, emtsv/emTag)
3. formátumátalakítás (**saját szkript**)
4. kézi ellenőrzés (Google Spreadsheets)
5. formátumátalakítás (**saját szkript**)
6. elemzés (emtsv/emDep)
7. formátumátalakítás (emtsv/emCoNLL)
8. kézi ellenőrzés (WebAnno)

[8] Mivel az annotálási munka jelenleg is folyik, az aktuális méretet lásd a korpusz repozitóriumában.

[9] `https://hu.globalvoices.org`

9. zéró létigék és igei ellipszisek kézi beillesztése (szövegszerkesztő)
10. zéró névmások beillesztése (**saját szkript**)
11. automatikus névmási anaforafeloldás (**saját szkript**)
12. kézi ellenőrzés és koreferenciaannotálás (Google Spreadsheets)
13. formátumátalakítás (**saját szkript**)

A kézi ellenőrzést igénylő munkafázisok esetében az annotátorok munkaidő-nyilvántartást vezettek, ahol nem csak azt rögzítették, hogy mely fájlokkal végeztek, hanem azt is, hogy az adott fájl ellenőrzésekor milyen nehézségekbe ütköztek. Ezen kívül azt is követtük, hogy az egyes fájlok ellenőrzése – az egyes elemzési szinteken – mennyi időt vett igénybe, ezáltal a korpusz további bővítésének költségei is kalkulálhatóak. Az 1. táblázat azt mutatja, hogy átlagosan hány percet vett igénybe egy dokumentum ellenőrzése a különböző elemzési szinteken.

	perc/dokumentum
morfológiai egyértelműsítés ellenőrzése	0:24:13
függőségi elemzés ellenőrzése	0:29:23
anaforák beillesztésének ellenőrzése	0:34:22

1. táblázat. A kézi ellenőrzéshez szükséges idő a különböző elemzési lépések után.

Érdemes szem előtt tartani a tényt, hogy az első néhány fájl ellenőrzése mindig több időt vett igénybe. Az annotátorok minden felmerülő problémát, nehézséget jeleztek, így az annotálási útmutató is finomodott, egyre pontosabb és világosabb iránymutatást biztosított, így a munka is felgyorsult.

A folyamatok ellenőrzéshez egy összevető program is készült, az emDiff[10], amely lehetővé teszi az eltérő tokenizálású szövegek simítását[11] és az egyes oszlopok tartalmának összevetését. Ennek köszönhetően nem csak a több annotátor által annotált ugyanazon szövegek összevetésére alkalmas, hanem annotátorok közötti egyetértés számítására[12] is. Végül az annotátorok által ellenőrzött végleges verzió és az egyes elemzők által produkált kimenetek is összevethetőek, így ezeknek az elemzőknek a teljesítménye is kiértékelhető a program segítségével. A program az `emtsv` moduljaként is futtatható.

Az egyes lépések között a fájlok formátuma többször is változik, ahol a rákövetkező lépés bemeneti fájlformátuma eltér. A folyamat legutolsó lépése a fájlok átalakítása az `emtsv` által használt rugalmas formátumra.

A következőkben részletezzük a korpusz építésének egyes lépéseit.

[10] `https://github.com/vadno/emdiff`

[11] a Python `difflib` csomagjával (`https://docs.python.org/3/library/difflib.html`)

[12] az `nltk. metrics` csomaggal (`https://www.nltk.org/api/nltk.metrics.html`)

4.1. Szöveggyűjtés és előkészítés

A fent említett forrásokból több mondatot tartalmazó szövegeket gyűjtöttünk, hiszen az anafora- és koreferenciaviszonyok mondathatárokon is átívelnek. A szövegek hossza 5 és 27 mondat között, a mondatok hossza 3 és 71 token között van (az írásjeleket külön tokennek számolva).

Az összegyűjtött szövegeket `emtsv` elemzőeszköz számára megfelelően kellett előkészíteni. Bár a pilotkorpusz szövegei standard helyesírásúak voltak, szükségesnek tartottunk minden szöveget átnézni. A Wikipédia és a Global Voices szövegeiben is bőven találtunk nem standard helyesírású szöveget, ezeket a nyers szövegek átolvasása során kézzel javítottuk. A szövegeket egyszerű szövegfájlokban (`txt` kiterjesztéssel, UTF-8 karakterkódolással) tároltuk el, szövegenként külön fájlban.

4.2. emToken, emMorph, emTag

Az `emtsv` megfelelő moduljai (emToken (Mittelholcz, 2017), emMorph (Novák, 2014; Novák és mtsai, 2016; Novák, 2003) és emTag (Orosz és Novák, 2012, 2013)) segítségével történő elemzés kimenete egy négy oszlopot tartalmazó fájl, aminek a formátumát röviden már a 3. fejezet ismertette. Az oszlopok tartalma a következő: token, emMorph kimenet, egyértelműsített tő, egyértelműsített morfológiai címke. Az emMorph kimenet a szó összes lehetséges elemzését – különböző címkeformátumokban – és az azokhoz tartozó tövet tartalmazza. Az egyértelműsített morfológiai címke az emMorph morfológiai címkekészletét használja.

4.3. Kézi ellenőrzés

Az első kézi ellenőrzés a tokenizálás, a tövesítés és az egyértelműsített morfológiai címke ellenőrzését jelentette. A feladat elvégzéséhez a be- és kimeneti formátum rugalmassága, az ergonomikus és könnyen használható felület, könnyű elérés, a verziókövetés és a kollaboratív felület kritériumainak a Google Spreadsheets felelt meg.

A három nyelvész annotátor az előkészített (összegyűjtött, átnézett, `emtsv`-vel tokenizált, morfológiailag elemzett és egyértelműsített), valamint a Google Spreadsheets formátumára igazított szövegeket szerkesztette. A feltételes formázások célja, hogy vezessék az annotátor szemét a munka során, kiemeljék a potenciálisan javítandó elemeket, valamint visszacsatolást nyújtsanak a javításról.

Az annotátorok a táblázatban a tokenizálás és a tövesítés javítása mellett az egyértelműsített morfológiai címkét (tehát az emTag kimenetét) javították. Ehhez az emMorph által produkált összes lehetséges morfológiai elemzés rendelkezésre állt, amelyek közül ki kellett választani a helyes elemzést a tővel együtt. Ha a morfológiai elemzések közül egyik sem volt helyes, akkor kézzel is meg lehetett adni a megfelelő elemzést.

A tokenizálási hibák javítására kézzel beírható parancsokat határoztunk meg, amelyeket aztán az exportált `csv` feldolgozásakor automatikusan értelmez egy

szkript és azoknak megfelelően módosítja a tokenizálást (sort töröl vagy sort szúr be az annotátor által megadott tartalommal).

Az exportált `csv` feldolgozásakor a tokenizálási javításokra vonatkozó parancsok értelmezése mellett az `emtsv` formátumára történő visszaalakítás is megtörtént. A kézi javítás után a szöveg tehát pontosan ugyanúgy néz ki, mint javítás előtt, különbség a kijavított mezőkben van csupán.

A szövegek egy részét az összes annotátor ellenőrizte, így ezeken a szövegeken kiszámolható az annotátorok közötti egyetértés. A javítás során az annotátorok arra vonatkozóan nem kaptak útmutatót, hogy az emMorph címke által megjelenített derivációkat és a szóösszetételeket hogyan kezeljék, hiszen a függőségi elemző már egy csak az inflexiós jegyeket megjelenítő címkekészlet alapján dolgozik (a részleteket lásd a 4.4. alfejezetben). Hogy az ebből fakadó különbségeket ne vegyük figyelembe az annotátorok közötti egyetértés számolásakor, már az emMorph-ról erre az inflexiós jegyeket tartalmazó készletre konvertált címkéket használtuk. A 4 315 tokennyi, mindhárom annotátor által ellenőrzött szövegre kapott eredmény Krippendorff alfában (Artstein és Poesio, 2008) kifejezve: 0,976.

Az annotációs útmutató és az ahhoz tartozó kiegészítő dokumentum, a feltételes formázásokat tartalmazó Google táblázat, valamint a táblázat formátumára alakító szkript elérhető a korpusz repozitóriumában.

4.4. emDep

A címkék kézi javítása után a szövegek szintaktikai elemzését az `emtsv` függőségi elemző modulja végez el. A függőségi elemzés előtt az emMorph címkét át kell alakítani a függőségi elemző modul számára emészthető UD-kompatibilis morfológiai címkére, amely megegyezik a Szeged Dependency Treebank (Vincze és mtsai, 2010) és az `emtsv` függőségi elemzője[13] (Zsibrita és mtsai, 2013), az emDep címkekészletével[14]. A továbbiakban erre a címkekészletre UD-ként hivatkozunk[15]. Ezt a konverziót is az `emtsv` egy modulja[16] végzi el. Az UD-re való konvertáláskor a derivációra vonatkozó egyes információk elvesznek. Azzal, hogy az emMorph címkék lettek kézzel javítva, és nem a már UD-re konvertált címke, többletmunkát végeztünk. Ezzel a többletmunkával azonban elértük, hogy a korpusz ezen rétege, a morfológiai egyértelműsítés is kézzel ellenőrzött minőségű.

A következő lépésben a javított és az `emtsv` formátumának megfelelően visszaalakított szövegeket az `emtsv` függőségi elemzőjével elemeztük.

4.5. Kézi ellenőrzés és zéró létigék

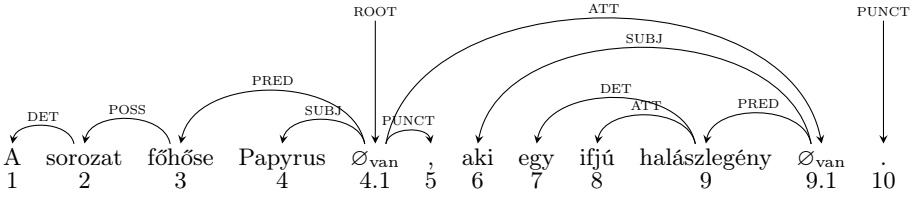
A javításhoz a WebAnno[17] általános célú, webalapú eszközt (Eckart de Castilho és mtsai, 2016) használtuk, hiszen a legtöbb fontos szempontnak megfelelt.

[13] `http://e-magyar.hu/hu/textmodules/emdep`
[14] A címkekészleteket Vadász és Simon (2019) részletesen ismerteti.
[15] `https://github.com/dlt-rilmta/panmorph/blob/master/panmorph_ud.pdf`
[16] `https://github.com/vadno/emmorph2ud`
[17] `https://webanno.github.io/webanno`

Drag-and-drop módszerrel használható, az elemzés különböző fázisaiban lévő dokumentumok is feltölthetőek, nem csak annotálásra, hanem javításra is használható. Olyan kiegészítő funkciókkal is bír, mint a több annotátor által kezelt dokumentumok összevetése, a munka egyszerű nyomonkövetése és az annotátorok közötti egyetértés különböző mérőszámok alapján történő automatikus kiszámolása. Az eszköz rugalmasnak mondható, hiszen saját elemzési rétegeket is megfogalmazhatunk. A WebAnno egy szerveren fut, az annotátorok pedig a megszokott böngészőjükön keresztül használhatják. A függőségi elemzés után az `emtsv` moduljaként működő konverterrel[18] alakítottuk át a kimenetet a WebAnno számára emészthető CoNLL-U formátumra. Az ellenőrzést három nyelvész annotátor végezte.

Használat közben azonban mégis felmerültek problémák ezzel az eszközzel kapcsolatban. Bár a tokenizálási hibák javítására már korábban volt lehetőség, mégis előfordult, hogy a függőségi elemzés kimenetének javításakor találkoztunk ilyen hibákkal. Sajnos a WebAnno felületén token törlésére vagy beszúrására nincs mód, így ezt a problémát egy utófeldolgozó lépésben kellett kezelni.

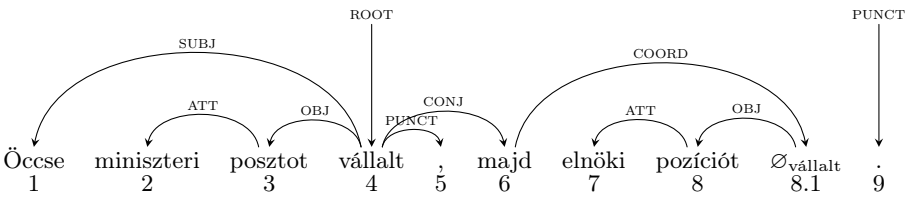
Az esetleges tokenizálási hibák javításával egyidőben a zéró létigék és az elliptált igék beszúrása is megtörtént. A zéró létigék beszúrását kézzel végeztük el azokban a mondatokban, amelyekben nem volt finit ige. A zéró létigék új tokenként kerülnek a fájlba arra a helyre, ahol múlt időben testes létigeként jelennének meg, saját kombinált indexet kapnak, ami a zéró létigét megelőző elem ID-jéből képződik. Az 1. példában egy olyan mondat látható, ahol a függőségi fába két zéró létigét is be kellett illeszteni.



1. ábra: Az összetett mondat fölérendelt tagmondatának zéró létigéje alá van rendelve az alárendelt mellékmondat zéró létigéje. A második sorban a kiosztott indexek láthatók a zéró létigék kombinált indexével együtt.

Az igei ellipsziseket is jelöltük a korpuszban, hiszen gyakran találkoztunk olyan tagmondatokkal, amelyekben az elliptált ige hiánya miatt nem lehetett megfelelő anyacsomóponthoz kötni az egyes bővítményeket. A zéró létigékhez hasonlóan kézzel illesztettük a mondatfába az elliptált finit igéket. Az elliptált ige a zéró létigéhez hasonló, kombinált indexet kapott. A 2. példában egy olyan mondat látható, ahol a függőségi fába egy elliptált igét kellett beilleszteni.

[18] `https://github.com/vadno/emconll`



2. ábra: A mellérendelés első mondatában szereplő ige a második mondatban testetlenül van jelen. Ezért egy zéró alakot illesztettünk be a függőségi fába, így a második mondatban szereplő vonzat már kapcsolódni tud a saját testetlen igéjéhez.

A korpusz 463 beillesztett zéró létigét és 25 beillesztett elliptált igét tartalmaz.

4.6. A zérónévmások beillesztése

A zérónévmásokat egy saját szkript, az emZero[19] illeszti be, amelynek bemenete a tokenizált és (javított) tövesítéssel, morfológiai egyértelműsítéssel és függőségi elemzéssel ellátott szöveg. Egyszerű szabályok mentén végzi az elemek beillesztését és a szabályok alkalmazása során különböző elemzési rétegek tartalmára támaszkodik (tő, morfológiai címke, függőségi elemzés).

A program a következő helyekre illeszt be zérónévmást:

- finit ige alanyának, ha annak nem volt testes alanya
- határozott ragozású finit ige tárgyának, ha annak nem volt testes tárgya
- birtok birtokosának, ha annak nem volt testes birtokosa
- ragozott és ragozatlan infinitívusz alanyának

A zérónévmások beillesztése után a mondatfában plusz ágak jelennek meg. A zéró elemek is saját ID-t kapnak, a tsv-be pedig az alany az ige után, a tárgy az ige (és a zéró alany) után, a birtokos pedig a birtok után kerül és egy kombinált ID-t kap, ami az őt megelőző elem ID-jéből és a zéró elem szintaktikai szerepének rövidítéséből (SUBJ, OBJ, POSS) áll. A zéró elemek szófaja névmás (PRON), a morfológiai jegyeik között pedig az ige vagy a birtok alapján kiszámolható szám és személy jegyek jelenhetnek meg.

A program az `emtsv` moduljaként is futtatható.

4.7. Névmási anaforák beillesztése

A következő lépésben a névmási anaforikus kapcsolatokat is egy szabályalapú szkript szúrja be. A program megkeresi a névmásokat, majd a mondatban szereplő többi szó szófaji, morfológiai és szintaktikai információira támaszkodva egyszerű szabályok alapján dönt.

[19] `https://github.com/vadno/emzero`

A szkript jelenleg csak a személyes névmások előzményét keresi meg, a többi típust kézzel kell beilleszteni. A személyes névmások előzménykeresésének egyszerű algoritmusát Pléh és Radics (1976) alapján dolgoztuk ki. Az algoritmus az alany antecedensének keresésekor például az alábbihoz hasonló szabályok alapján dönt:

1. ha az ige alanya zéró névmás és az ige ragozása az előző mondat igéjének ragozásához képest nem változott, akkor az alany antecedense az előző mondat igéjének alanya
2. ha az ige alanya mutatónévmás, akkor annak antecedense az előző mondat nem alanyi argumentuma

4.8. Kézi ellenőrzés és koreferenciaannotálás

Az automatikusan beillesztett zéró névmások és névmási anaforák ellenőrzését, valamint a koreferenciaannotálást négy nyelvész annotátor végezte.

Számos annotációs eszköz található, amelyek segítségével lehet anafora- és koreferenciaéleket annotálni a szövegekben (pl. WebAnno, brat[20] (Stenetorp és mtsai, 2012), TrEd[21] stb.). Vannak olyan eszközök is, amelyek az annotáció javítására használhatók és például CoNLL-U formátumban képesek feldolgozni az adatot. Legjobb tudomásunk szerint olyan eszköz azonban nem áll rendelkezésre, amely minden fontos kritériumunknak megfelelt volna, és emellett a zéró elemek kezelésére is alkalmas lenne (mert például a CoNLL-U formátum a hivatalos formátumleírás[22] alapján nem teszi lehetővé, hogy zéró elemek szerepeljenek a függőségi fában).

Az automatikusan beillesztett zérónévmások és anaforikus kapcsolatok ellenőrzését, valamint a koreferenciakapcsolatok beillesztését így ismét feltételes formázásokkal ellátott Google Spreadsheets táblázatokban végeztük el. Az anaforikus- és koreferenciakapcsolatokat két oszlopban kellett jelölniük az annotátoroknak, egyikben annak az elemnek az ID számát kellett megadni, amellyel a visszautaló elem kapcsolatban áll, a másikban pedig a kapcsolat típusát. A korpuszban az alábbi anaforikus kapcsolattípusokat jelöltük (zárójelben a korpuszban szereplő jelölésükkel):

- személyes (**prs**)
- mutató (**dem**)
- kölcsönös (**recip**)
- visszaható (**refl**)
- vonatkozó (**rel**)
- birtokos (**poss**)

Az automatikus névmási anaforákat beillesztő program a személyes névmások előzményén kívül nem ad számot a többi névmásról és azokról a kapcsolatokról,

[20] `http://brat.nlplab.org`
[21] `http://ufal.mff.cuni.cz/tred`
[22] `https://universaldependencies.org/format.html`

amelyekben ezek szerepelnek. Ilyen az általános alany szerepben álló zéró névmás, amelynek referenciája nehezen megragadható (**vastag betűvel** kiemelve az általános referenciájú alannyal rendelkező igéket (1. példák).

(1) a. ... a Kínai Kommunista Párt egyik volt vezetője, akit hazaárulás miatt **elítéltek** ...

b. 1883-ban **említették** először az orthodox hitközségnek adományozott területként ...

Hasonlóak azok esetek, amikor a szöveg írója megszólítja az olvasót (2. példa, ahol **vastag betűvel** kiemeltük azokat az igéket, amelyeknek alanya az író vagy az olvasó). Ez a típus nem gyakori a hírszövegekben vagy a Wikipédia-szövegekben, ugyanakkor a korpusz más műfajú szövegekkel (pl. szépirodalom, személyes szövegek) történő bővítésénél gyakrabban előfordulhat.

(2) A születésnapi ajándékoknak is nagyon **örülünk**, ha **szeretnéd** támogatni a munkánkat, **küldj nekünk** adományt, vagy **vegyél** egyet az NSA-s karácsonyi **üdvözlőlapjaink** közül, amelyet a Creative Time-nál dolgozó **barátaink** terveztek.

Új címkéket vezettünk be ezekre az esetre: az **arb** az általános alany, az **addr** a címzett, a **speak** a beszélő/író referenciáját jelöli. Azzal, hogy bevezetjük a szöveg szereplői közé a beszélőt és a címzettet, jelölni tudjuk, ha a szövegben szereplő névmások ezen szereplők valamelyikével koreferens antecedensre utalnak vissza.

Az egyes koreferencakorpuszokban, így a SzegedKoref annotációjában is különböző típusú koreferenciakapcsolatokat (pl. ismétlés, variáció, szinonima, hipernima, hiponima és holonima stb.) jelölnek. Az annotálás tervezése során és a szövegeket megfigyelve azonban számos nehézségbe ütköztünk ezekkel a típusokkal kapcsolatban. A korpusz annotációja így csupán kétféle koreferenciatípust különböztet meg.

A **coref** címkével jelölt koreferenciatípus magában foglalja az összes olyan koreferenciakapcsolatot, amely két azonos referenciájú elemet köt össze, így például az ismétlést, a szinonimát, hiper- és hiponimát. A **holo** címkével jelölt kapcsolattípus pedig azt jelenti, hogy a két szó referenciája között rész-egész viszony áll fenn, pontosabban a második szó referenciája része az első szóénak.

Míg a koreferenciakapcsolatok előzménye (amellyel közös referenciájuk van vagy referenciájuk között rész-egész viszony áll fenn) mindig testes szó, addig a (testes vagy zéró) névmások előzménye (antecedense) lehet tartalmas szó, illetve testes vagy testetlen névmás. Ennek megfelelően az anaforikus és koreferenciakapcsolatok nem folyamatos láncot képeznek a szövegen át, hanem elágazásokat, kitérőket is tartalmaznak.

A 2. táblázat összefoglalja, hogy a korpusz összesen hány visszautalást tartalmaz az egyes kapcsolattípusokból. Mindemellett az ellenőrzés végén a korpusz 2 346 zéró alanyt, 260 zéró tárgyat és 914 zéró birtokost tartalmaz.

kapcsolat	előfordulás
prs	1 497
dem	147
recip	11
refl	18
rel	447
poss	0
arb	316
speak	5
addr	1
coref	1 582
holo	202

2. táblázat. Az anaforikus és koreferenciakapcsolatok előfordulása a KorKorpuszban.

4.9. Nehézségek

A koreferencia annotálásakor számos nehézséggel találtuk szembe magunkat, amelyek kezelésére a szakirodalom sem tudott megnyugtató választ kínálni. A 3. példa azt a problémát illusztrálja, amikor a referens állapota megváltozik (itt: meghal). Vajon a holttest koreferens az emberrel?

(3) Három hónap telt el az **újságíró házaspár**, Sagar Sarwar és felesége, Meherun Runi meggyilkolása óta. A **holttesteket** már exhumálták is, hogy megismételjék a boncolást.

A 4. példa azt a nehézséget szemlélteti, amikor egy szó előzménye, amellyel koreferens, több tagból áll.

(4) **Papyrus** bátor és megmenti **Thèti-Chèri-t**. **A két egymásra lelt barát** küldetést kap az istenektől, hogy védelmezzék meg a fáraót.

A két egymásra lelt barát egyszerre koreferens *Papyrussal* és *Thèti-Chèrivel*, sőt, csak az egyikükhöz kötni feltétlenül hibás volna. Ugyanakkor a jelenlegi annotációs séma alapján csak egyetlen előzményhez lehet kötni. Az se sokat javítana a helyzeten, ha mellérendelés állna fenn *Papyrus* és *Thèti-Chèri* között. Ugyan ebben az esetben már ábrázolható lenne a koreferenciakapcsolat *a két egymásra lelt barát* és a mellérendelés feje között, viszont a feloldása többértelmű lenne, hiszen nem lehetne eldönteni, hogy a teljes mellérendelő szerkezetre, vagy csak a fejében lévő elemre történt-e a visszautalás.

Az ezekhez hasonló problémás esetek kezelésére külön irányelveket kell kidolgozni, ám az ezekkel kapcsolatos döntések még előttünk állnak.

5. A pilotkorpusz hasznosíthatósága

A KorKorpusz már a jelenlegi pilot fázisban is több célra hasznosítható. A kézzel annotált, jó minőségű adat értékét nem lehet eléggé hangsúlyozni, legyen szó bármely elemzési feladatról. A KorKorpusz kézzel javított elemzési rétegei lehetővé teszik, hogy megvizsgáljuk, hogy a korpuszépítéshez használt egyes eszközök milyen minőségű elemzést biztosítottak, így nemcsak a KorKorpusz építéséhez újonnan készült előelemző eszközök, hanem az `emtsv`-ben használt modulok teljesítménye is kimérhető.

Az emTag (Orosz és Novák, 2012, 2013) teljesítményét a 4. fejezetben ismertetett emDiff segítségével vizsgáltuk meg azon a 122 fájlon, amelyekben az egyértelműsítés eredményét kézzel javítottuk. Összevetettük az egyértelműsített tő és az egyértelműsített morfológiai címke mezők tartalmát. Az eredményeket a 3. táblázat tartalmazza.

	pontosság
tő	98,15%
morfológiai címke	95,40%

3. táblázat. Az emTag teljesítménye pontosságban (*accuracy*) kifejezve.

Az eredmény alapján elmondható, hogy kevés esetben kellett kézzel kivajítani a címkéket. Az egyes hibatípusok csoportosítása a későbbiekben segítséget nyújthat az emTag kimenetének automatikus javításában is.

6. További tervek

A KorKorpusz további fejlesztésének két iránya van: egyrészt a rendelkezésre álló eszközök és dokumentációk segítségével a korpusz további szövegekkel való kibővítésével, másrészt az egyes munkafolyamatok még könnyebbé és ergonomikusabbá tételével. Távlati tervek között szerepel az elkészített eszközök további javítása, valamint a koreferenciakapcsolatok automatikus beillesztésének kidolgozása.

Köszönetnyilvánítás

Hálámat fejezem ki az annotátoraimnak, Bencze Norbertnek, Bognár Ivettnek, Fegyó Kingának és Fodor Grétának, akik a monoton feladatok elvégzése mellett friss ötleteikkel folyamatosan inspiráltak.

Hivatkozások

Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34(4), 555–596 (2008)

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C.: A web-based tool for the integrated annotation of semantic and syntactic structures. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). pp. 76–84. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), `https://www.aclweb.org/anthology/W16-4011`

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the 8th International Conference, TSD 2005. pp. 123–131. Springer, Karlovy Vary, Czech Republic (2005)

van Deemter, K., Kibble, R.: What is coreference, and what should coreference annotation be? In: Coreference and Its Applications. pp. 90–96 (1999)

Erjavec, T.: MULTEXT-East Morphosyntactic Specifications. Version 3.0 (May 2004), http://nl.ijs.si/ME/Vault/V3/msd/html/

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – egy formátum mind felett. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2019)

Miháltz, M.: Tudásalapú koreferencia- és birtokviszony-feloldás magyar szövegekben. Általános Nyelvészeti Tanulmányok 24, 151–166 (2012)

Miháltz, M., Naszódi, M., Vajda, P., Varasdi, K.: NP-koreferenciák feloldása magyar szövegekben a magyar wordnet ontológia segítségével. In: Tanács, A., Csendes, D. (szerk.) V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016). pp. 138–146. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2007)

Mittelholcz, I.: `emToken`: Unicode-képes tokenizáló magyar nyelvre. In: Vincze, V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). pp. 70–78. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2017)

Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)

Novák, A.: A new form of Humor – Mapping constraint-based computational morphologies to a finite-state representation. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)

Novák, A., Rebrus, P., Ludányi, Zs.: Az `emMorph` morfológiai elemző annotációs formalizmusa. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). p. (jelen kötetben). Szeged (2017)

Novák, A., Siklósi, B., Oravecz, Cs.: A new integrated open-source morphological analyzer for Hungarian. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk,

J., Piperidis, S. (szerk.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)

Orosz, Gy., Novák, A.: PurePos 2.0 – an open source morphological disambiguator. In: Sharp, B., Zock, M. (szerk.) Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science. pp. 53–63. Wroclaw (2012)

Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. pp. 539–545. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (Sep 2013), https://www.aclweb.org/anthology/R13-1071

Pléh, Cs., Radics, K.: „Hiányos mondat", pronominalizáció és a szöveg. Általános Nyelvészeti Tanulmányok 11(1), 261–277 (1976)

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: A web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)

Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)

Vadász, N., Simon, E.: Konverterek magyar morfológiai címkekészletek között. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 99–112. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2019)

Vincze, V., Hegedűs, K., Farkas, R.: SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 312–322. SZTE TTIK Informatikai Tanszékcsoport (2015)

Vincze, V., Hegedűs, K., Sliz-Nagy, A., Farkas, R.: SzegedKoref: A Hungarian coreference corpus. In: Proceedings of the 11th Language Resources and Evaluation Conference. European Language Resource Association, Miyazaki, Japan (2018)

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (szerk.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)

Automatikus tematikuscímke-ajánló rendszer sajtószövegekhez

Yang Zijian Győző[1,2], Novák Attila[1,2], Laki László János[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
{yang.zijian.gyozo, novak.attila, laki.laszlo}@itk.ppke.hu

Kivonat Cikkünkben sajtószövegek automatikus tematikus címkézésével kapcsolatos kutatásunk eredményét, illetve a kutatás keretében létrehozott automatikus címkézőrendszert mutatjuk be. A rendszerhez olyan felhasználói felületet hoztunk létre, amely lehetővé teszi a felhasználó számára a rendszer bizonyos paramétereinek beállítását. Ennek segítségével az ajánlott kulcsszólista fedése és pontossága testre szabható. Bemutatjuk a különböző modellparaméterek beállításának hatását a címkézés teljesítményére.
Kulcsszavak: címkeajánlás, kulcsszavazás, fastText, SentencePiece tokenizáló

1. Bevezetés

A webes hírportálokon megjelenő szövegeket gyakran különböző tematikus címkékkel látják el. Ezek szerepe többféle. Egyrészt lehetővé teszik a látogatók számára, hogy kifejezetten egy-egy számukra érdekes témával, személlyel, eszközzel stb. kapcsolatos cikkeket vagy egyéb tartalmakat a kulcsszavakra (tematikus címkékre) szűrve megjelenítsék. Másrészt a kulcsszavakat az adott cikkhez kapcsolódó egyéb cikkek vagy tartalmak szűréséhez/megjelenítéséhez is használják. Emellett a tematikus kulcsszavak arra is használhatóak, hogy az üzemeltető regisztrált vagy egyéb módon nyilvántartott felhasználói számára az érdeklődésüknek megfelelő testre szabott tartalmakat ajánljon.

Szerepet játszanak a címkék a keresőmotorok (pl. a Google) találatrangsorolási algoritmusaiban is. A html tartalom megfelelő kulcsszó-metacímkéinek tartalmaként megadott kulcsszavakat a keresőmotorok korábban egyértelműen előrébb rangsorolták, mint a puszta szövegszótalálatokatA forgalomnövelés céljából végrehajtott manipulatív keresőoptimalizálás céljából bevezetett sok hamis címke megjelenése miatt a keresőmotorok üzemeltetői később csökkentették vagy mellőzni kezdték a kulcsszó-metacímkék tartalmának figyelembevételét a találatok rangsorolásánál.

Ennek ellenére a korábban felsorolt okokból, illetve mert a valóban releváns kulcsszavak továbbra is fontosak, illetve a kulcsszavak köré rendezett tematikus cikkgyűjteményoldalakat a keresőmotorok továbbra is lelkesen indexelik (itt a

cím, illetve az url része az adott kulcsszó), a megfelelő kulcsszavak cikkekhez rendelése továbbra is fontos az online sajtó számára. Ugyanakkor bár az adott szöveghez kapcsolódó tematikus kulcsszavak automatikus hozzárendelésére számos algoritmikus megoldás létezik, sok online is megjelenő szövegarchívumban a kulcsszavak tartalomhoz való hozzárendelését jelenleg is kizárólag emberi munkával végzik.

Van, ahol központilag egyetlen – tipikusan könyvtáros végzettségű – munkatárs végzi a cikkek kulcsszavazását. Ez a megoldás viszonylag egységes és jól átgondolt címkehasználatot eredményez, azonban a hosszú átfutási idő miatt – a címkézőnek el kell olvasnia minden cikket – csak viszonylag korlátozott mennyiségű tartalom címkézése oldható meg így. Ez a módszer egy hetilap esetében alkalmazható, azonban egy adott idő alatt jóval nagyobb mennyiségű – ráadásul online karbantartott, illetve időnként módosított – tartalmat generáló webes tartalomszolgáltató, illetve hírportál esetében nincs idő arra, hogy egyetlen dedikált személy végezze a címkézést. Ebben az esetben a manuális címkézést maguk a szerzők végzik, és az egységesség irányába csak a szerkesztőségi irányelvek, illetve esetlegesen a szerkesztőség által használt tartalomkezelő rendszerbe (CMS) integrált prediktív keresésen alapuló ajánló használata mutat (ahogy a szerző elkezdi gépelni a kulcsszót, a korábban már használt egyező kezdetű kulcsszavak listája megjelenik, és a lista tovább gépelve egyre szűkül). Ugyanakkor egyrészt minden szerző bármikor új kulcsszót vehet fel, másrészt a prediktív ajánló használata egy félregépelt kulcsszó gyakori használatához is vezethet.

Cikkünkben egy olyan tartalomkezelő rendszerekbe integrálható rendszert mutatunk be, amely automatikus kulcsszóajánlással segíti a szerkesztőség munkáját. A korábban manuálisan címkézett szövegeken betanított modellt használunk az újonnan születő szövegekhez alkalmazható címkék megjósolására.

2. Kapcsolódó irodalom

Magyar szövegek tematikus kulcsszavazásával kapcsolatban fontos korábbi eredményről számol be Farkas (2009). Ebben a cikkben az [origo] hírportál korábbi kulcsszavakkal el nem látott tartalmainak teljesen automatikus címkézésére alkalmazott megoldást mutat be a szerző. Megoldásuk igen sokrétű: a cikkek szövegét különböző szintű elemzésnek (szófaji és névelemcímkézés) vetették alá, névszói csoportokat kerestek, azokat az egységes címkekészlet létrehozása érdekében normalizálták/lemmatizálták, emellett a címkéket különböző osztályokba (személy, földrajzi név, szervezet, egyéb téma) sorolták. A beszámoló ugyanakkor nem említi, hogy a rendszert az újonnan írt cikkek címkézésének segítésére használták volna. Az akkori fejlesztés során viszonylag kevés címkézett szöveg állt rendelkezésre, ezért elsősorban az adott szövegben szereplő kifejezések kiemelésére és normalizálására támaszkodtak a címkézéskor.

Mi abból indultunk ki, hogy nagyméretű manuálisan címkézett szöveghamaz áll rendelkezésre, ezért elsősorban ennek kiaknázására alapoztuk megoldásunkat, amelynek elsődleges célja a további alapvetően továbbra is emberi kontroll alatt végzett címkézési munka hatékony segítése. Választásunk a fastText programcso-

mag (Joulin és mtsai, 2017) címkézőalgoritmusára esett. Ebben a megoldásban a neurális osztályozóhálózat(ok) bemenetén az adott szöveg tokenjeinek, illetve token-n-eseinek reprezentációja jelenik meg (a bennük szereplő különböző hosszú karakter-n-gramok reprezentációjának eredőjeként), és az osztályozó ehhez a szövegreprezentációhoz és az egyes lehetséges címkékhez rendel illeszkedési értéket multinomiális logisztikus regresszió alkalmazásával. Megfelelő küszöbérték kiválasztásával az adott szövegre jól illeszkedő kulcsszavak elválaszthatóak a kevésbé jól illeszkedőektől. Bár megjelenése óta a fastText modellnél jobban teljesítő szövegosztályozó modellek is megjelentek (a cikk írásának idején az ilyen jellegű feladatokban az XLNet architektúra adja a legjobb teljesítményt nyújtó megoldást több angol nyelvű adatbázison (Yang és mtsai, 2019)), ezeknek komplexitása, hardver- és futásiidő-igénye a pontosságbeli teljesítménykülönbséget jóval meghaladó mértékben nagyobb, mint a fastTexté.

A fastText által előállított szóbeágyazási modell osztályozási feladatra való alkalmazását mutatja be Szántó és mtsai (2017), azonban az ott bemutatott osztályozási feladat egyszerű kétosztályos osztályozást jelentett (sport/videojáték) szemben az általunk kitűzött céllal, ahol sok ezer, sőt akár sok tízezer lehetséges egyedi címke közül kell kiválasztani az adott szövegre legjobban illeszkedő címkéket (amelyek száma a témától és a szöveg hosszától függően változhat).

3. A címkézőrendszer

Ebben a részben áttekintjük a címkézőrendszer architektúráját, illetve bemutatjuk röviden a mögöttes adatbázist.

3.1. A címkézőrendszer architektúrája

Az elkészült címkézőrendszer[1] egy REST alapú webes applikáció, ami két részből áll: frontend és backend. A frontend egy javascript bootstrap alapú felület, amin a felhasználó egy formon keresztül be tudja vinni az adatokat (cikk szövege, lead, cím, szerző, évszám), majd a *Címkézés* gombbal le tudja kérni a címkéző által javasolt címkéket, melyeket a felület megjelenít. A felhasználónak lehetősége van kiválasztani, hogy melyik modellel szeretné címkézni a cikket.

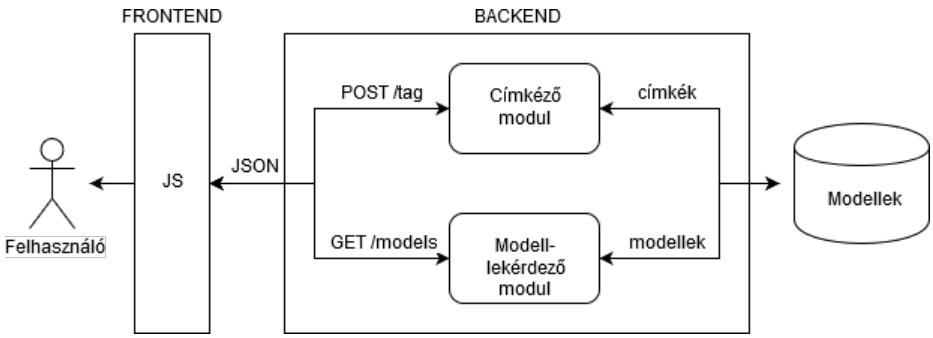
A megjelenített címkéknek három kategóriája van: kulcsszavak, tulajdonnevek és trendcímkék. A trendcímkék olyan címkék, amelyek valamilyen konkrét egyedi eseményre (egy konkrét választás, fesztivál, olimpia, konferencia stb.) utalnak, és inkább rövid távon van jelentőségük és értelmük. Ilyen például az *Oscar-gála 2018,* amely ugyan meglehetősen hasonló bármely más adott évben lezajlott Oscar-gálához, azonban az adott díjátadó időpontját övező viszonylag behatárolt időszakon kívül feltehetőleg nem hasznos címke. A kulcsszavak és tulajdonnevek statikus címkék, nem időszakhoz vagy konkrét eseményhez kötöttek. Ilyen például a *politika*, a *zene* stb., illetve a személyt, szervezetet stb. jelölő névcímkék mellett az általánosabb eseménytípust jelölő név jellegű címkék, mint pl. a dátum nélküli *Oscar-gála*.

[1] http://nlpg.itk.ppke.hu/projects/tagger

A felületen továbbá lehetőség van egy 0-tól 1-ig terjedő skálán beállítani, hogy milyen valószínűségű/konfidenciaértékű címkéket jelenítsen meg, valamint van egy „minimum 3” funkció, amelynek bekapcsolása esetén, függetlenül a konfidenciaküszöbtől, mindenképpen minimum 3 darab címkét jelenít meg.

A tesztanyagunknak készítettünk egy külön demófelületet, amelyre a címkézőfelületről át lehet navigálni. A tesztanyaghoz tartozó felület megegyezik a demófelülettel, annyi kiegészítéssel, hogy láthatóak a tesztkorpuszban szereplő szövegekhez tartozó referenciacímkék is, valamint megjeleníti a pontosság és a fedés értékeit is.

A címkéző másik része a backend. A frontend AJAX kéréssel tudja beküldeni a cikket és lekérni a backendtől az ajánlott címkéket. A frontend és a backend egymás között JSON formátumú adatcsomagokkal kommunikálnak. A backend egy Python nyelven implementált Flask[2] webszerver. A webszerver indításkor betölti az előre betanított címkézőmodelleket. Külön modellt hoztunk létre a statikus, a trend- és a tulajdonnévcímkékre. Emellett az online adatbázis cikkeiből tanított és a nyomtatott és az online adatbázis összevont anyagán betanított modell is kipróbálható. A modelleket rendszeresen újratanítjuk, hogy a szerzők által újonnan felvitt címkék bekerüljenek a rendszerbe. A trendcímkemodell tanításakor csak az utolsó fél év anyagát használjuk. A régebbi szövegeknél a trendcímkéket azok általános ekvivalensére cseréljük le.



1. ábra: A címkézőrendszer architektúrája

3.2. A korpusz

Tanító és tesztkorpuszként a hvg.hu által rendelkezésünkre bocsátott nyomtatott és online hírlapból vett cikkeket használtunk fel. Első kísérleteinket az 1. táblázatban bemutatott korpuszon végeztük. Később bővebb anyaghoz jutottunk: az online cikkadatbázis kiegészült a 2012 és 2016 közötti anyaggal. Az utóbbi bővített korpuszon végzett kísérleteinket a 6. részben mutatjuk be.

[2] https://palletsprojects.com/p/flask

Kísérleteink során többféle tanítóanyagot hoztunk létre: a nyomtatott hetilap anyagából, az online cikkekből, illetve a kettőt ötvöző hibrid korpuszt. A három különböző tanítóanyagból további változatokat hoztunk létre. Kísérleteztünk kisbetűsítéssel, stopword-listában szereplő szavak törlésével, ezek kombinációjával, valamint a szövegeken betanított SentencePiece tokenizáló alkalmazásával.

	Nyomtatott hírlap	Online hírlap
Felépítés	id, cím, lead, cikk szövege, szerző kategória, év, dátum	id, cím, alcím, cikk szövege, szerző, kategória, létrehozás és módosítás dátuma
Címkefajták	kulcsszavak, személy, szervezet, földrajzi	kulcsszavak
Megjelenés	hetilap	napilap
Időszak	1994-2017	2017-2018
Témák	gazdaság, politika, tudomány, sport kultúra, pszichológia	gazdaság, politika, tudomány, sport kultúra, pszichológia, blog,
Statisztika	cikk: 119077 cikk címke: 1023 db token: ∼73 millió type: 62 ezer	cikk: 86256 cikk címke: 76654 db token: ∼35 millió type: ∼33 ezer

1. táblázat. Nyomtatott és online hírlap tulajdonságai

3.3. Használt eszközök

Modelljeinket a fastText programkönyvtár a SentencePiece tokenizáló használatával készítettük.

A fastText (Joulin és mtsai, 2017) egy nyílt forráskódú programkönyvtár osztályozási feladatra és szövegreprezentációs modell létrehozására. Az eszközt a Facebook fejlesztette C++ nyelven.

A SentencePiece nevű eszköz egy felügyelet nélküli szövegtokenizáló és detokenizáló. Implementálva van benne a BPE algoritmus, ami egy unigram nyelvmodellel (Kudo, 2018) van súlyozva. Használatával elhagyhatók a nyelvspecifikus előfeldolgozási lépések, mint például a tokenizálás vagy a kisbetűsítés. A módszer lényege, hogy a természetes szöveget úgy alakítja át, hogy abban a különböző tokenek száma egy paraméterként megadható korlátos szám legyen, ezért az így létrejött tanítóanyagban általában nem lesznek ismeretlen szavak. (A tanítóanyagban nem szereplő ismeretlen karakterek (pl. idegennyelvű szövegrészekből), s így ismeretlen tokenek ritkán előfordulhatnak.) Ennek köszönhetően a neurális modellek paraméterszáma nagymértékben csökkenthető a hagyományos szóalapú modellekhez képest. A módszer a neurális gépi fordítás területéről származik, és a szövegfeldolgozásra használt mélytanuló modellekben nagyon elterjedt a használata.

4. Kísérletek

Első lépésként az 1. táblázatban látható nyomtatott cikkadatbázissal (NYC) kísérleteztünk. Az alábbi modelleket hoztuk létre:

- NYC-T: tokenizált (T) szöveg
- NYC-TK: tokenizált és kisbetűsített (K) szöveg
- NYC-PKS: Sentence Piece tokenizált (P) és kisbetűsített szöveg; stopwords lista használta
- NYC-PS: Sentcence Piece tokenizált szöveg; stopwords lista használata
- NYC-P: Sentcence Piece tokenizált szöveg

Stopwords listához az NLTK[3] magyar nyelvű csomagjához tartozó stopwordlistát használtuk, amely 199 szót tartalmaz.

Következő lépésként az online cikkadatbázissal (OLC) kísérleteztünk. Az alábbi modelleket hoztuk létre:

- OLC-P: Sentence Piece tokenizált szöveg

5. Eredmények

A 2. táblázatban láthatóak az első lépés eredményei, amelyeket 0,8-as valószínűség mellett értünk el a nyomtatott cikkadatbázis anyagán. Az egyik szembetűnő eredmény, hogy a SentencePiece tokenizáló használata majdnem kétszeresére növelte a fedés értéket. A másik érdekes eredmény, hogy a legnagyobb pontosságot az a modell érte el, amelyik kizárólag SentencePiece tokenizálót használt, sem kisbetűsítést, sem stopwordlistát nem. De az is látható, hogy ha elhagyjuk sorban ezeket az eszközöket, ugyan a pontosság nő, a fedés értéke csökkenni kezd.

Modell	Pontosság	Fedés
NYC-T	0,749	0,152
NYC-TK	0,750	0,162
NYC-PKS	0,748	**0,362**
NYC-PS	0,768	0,300
NYC-P	**0,774**	0,284

2. táblázat. A nyomtatott cikkadatbázissal végzett kísérleteink

A 3. táblázatban látható néhány példa arra, hogy milyen címkéket ajánl a rendszerünk. Látható, hogy amikor a rendszer magas valószínűséggel becsül, azok a címkék majdnem azonosak az eredeti címkékkel. De az is látható, hogy az alacsonyabb pontossággal becsült címkék szintén elég közel állnak a témához, és általában jó ajánlások. Vannak esetek, mint például az „alkotmánybíróság”

[3] https://pythonspot.com/nltk-stop-words/

esetében, hogy csak egy címkét társítottak a cikkhez, ajánlórendszerünk pedig az összes hasonló jelentésű címkét visszaadta eredményül.

Van olyan eset is, amit a „belpolitika" példáján látható, hogy az eredeti címkék között nem szerepel az „önkormányzat", de a cikk tartalmát tekintve erősen összefügg ezzel, ezért az ajánlórendszerünk ezt gondolta legvalószínűbbnek.

A „fúzió" példában, bár a címkék között nem szerepel a sztrájk, de a cikk végén több mondat is szól a sztrájkról, ezért a rendszerünk ajánlja ezt a címkét.

Láthatunk továbbá példát arra, hogy az NYC modell az online cikkeire tesz ajánlást. Teljesítményt nem tudtunk mérni ezen, hiszen teljesen más a címkekészlet. De a példákból láthatjuk, hogy teljesen jól megközelíti a témát. A „stewardess" címke nem szerepel a NYC modellben, de helyette „foglalkoztatás" címkével egész jól közelíti.

6. Részletes kísérletek a címkegyakoriság és a teljesítménymutatók összefüggésével kapcsolatban

Az előző részekben említett előzetes kísérletek után részletes kísérleteket végeztünk a hetilapkorpuszon és a 2012–2018 közötti időszakból származó kibővített online anyagon. Szerettük volna megtudni, hogy milyen összefüggés van az egy-egy címkére a tanítóanyagban látott példák száma és a rendszer teljesítménye között. Mindkét korpuszt úgy osztottuk tanító- és tesztanyagra, hogy a korpuszban legalább 15-ször szereplő címkékre a tesztanyagban legyen legalább 5 példa.

A két korpuszból készített tanító és tesztanyagunk jellemzőit a 4. táblázatban foglaltuk össze. A hetilapkorpusz cikkei hosszabbak, így az online korpusz több mint háromszor annyi cikke szószámban kevesebb mint kétszer akkora terjedelmű. Bár a hetilapkorpusz jóval hosszabb időszakot ölel fel, a címkehasználat egységesebb: a másik korpusz 7,5-szer többféle címkét tartalmaz. Ami a névcímkék arányát illeti (a nagybetűt tartalmazó címkéket soroltuk ide): ezek teszik ki a címkeelőfordulások nagyjából 95%-át. Ugyanakkor a címketípusok (a különböző címkék) jóval nagyobb része név a hetilapkorpuszban (96% a tanítóanyagban), mint az online korpuszban (50% alatt). Ennek oka egyrészt a nem név jellegű tematikus címkék sokkal nagyobb változatossága, másrészt a ritkább névcímkék sokszor kicsit hanyag kisbetűs írásmódja. A név-fogalom homográf párok mindkét tagja (pl. *Bugyi–bugyi, Magyar Csapat–magyar csapat*) az esetek nagy részében összevonva kisbetűvel szerepel címketöbbértelműséget eredményezve. A tesztanyagban igen, de a tantóanyagban nem szereplő (OOV) címkelőfordulások aránya a hetilapkorpusz esetén 2,7%, az online korpusznál 6,9%.

A hetilapkorpusz anyagán tokenizálás nélkül, hagyományos tokenizálással, és a SentencePiece tokenizálóval is betanítottunk egy-egy modellt, az online anyagon csak az utóbbit teszteltük. A tanítás paraméterei azonosak voltak. Minden esetben one-to-many osztályozókat tanítottunk be a szövegenként változó címkeszám kezelésére, és 100 dimenziós vektorokat használtunk. Az online korpusz sokkal nagyobb elemszámú címkekészlete miatt az online modell betanítása jóval hosszabb időt igényelt annak ellenére, hogy a hetilapkorpusz 50 epochos tanításával szemben 30 epoch (teljes korpuszbejárás) volt a tanítás ideje.

Eredeti címkék	Ajánlott címkék
NYC modell: ajánlások a hetilapkorpuszból származó cikkekre	
fúzió, gazdaság, közlekedés légi közlekedés	légi közlekedés (0,999), fúzió (0,743) gazdaság (0,719), közlekedés (0,378) sztrájk (0,033), repülőterek (0,009) vállalatgazdaság (0,003)
alkotmánybíróság	jog (1,000), alkotmányjog (1,000) alkotmánybíróság (1,000), alkotmány (0,772) büntetőjog (0,699), jogtörténet (0,328) jogalkotás (0,197)
belpolitika, pártpolitika, tömegközlekedés	önkormányzat (0,994), belpolitika (0,971), pártpolitika (0,492), tömegközlekedés (0,133), politika (0,104), önkormányzati költségvetés (0,078), uniós támogatás (0,032)
NYC modell: ajánlások az online korpuszból származó cikkekre	
földrengés, irán földmozgás, utórengés	földrengés (1,000), természeti katasztrófa (0,999), katasztrófa (0,998), idősek (0,053) halálbüntetés (0,022), katasztrófavédelem (0,017)
repülés, stewardess, wizz air	légi közlekedés (1,000), repülőgépgyártás (0,160), foglalkoztatás (0,078), járműipar (0,038) repülőterek (0,010), gazdaság (0,002) közlekedés (0,002)
OLC modell: ajánlások az online korpuszból származó cikkekre	
földrengés, irán földmozgás, utórengés	földrengés (1,000), irán (0,979) örményország (0,508), utórengés (0,294) lövöldözés (0,192), halálos áldozat (0,178)
repülés, stewardess, wizz air	wizz air (1,000), légitársaság (0,818) repülés (0,803), lufthansa (0,314) repülőgép (0,307)

3. táblázat. Példák a címkézőrendszer ajánlásaira

korpusz		token				típus	
	cikk	szó	címke	név	OOV	címke	név
Hetilap tanító	94094	46,25M	0,46M	0,44M	-	24849	23822
Hetilap teszt	6902	3,33M	38775	37559	1036	5089	4152
Online tanító	328635	89,08M	1,33M	1,26M	-	186508	89711
Online teszt	45105	13,43M	0,21M	0,2M	14488	53568	24607

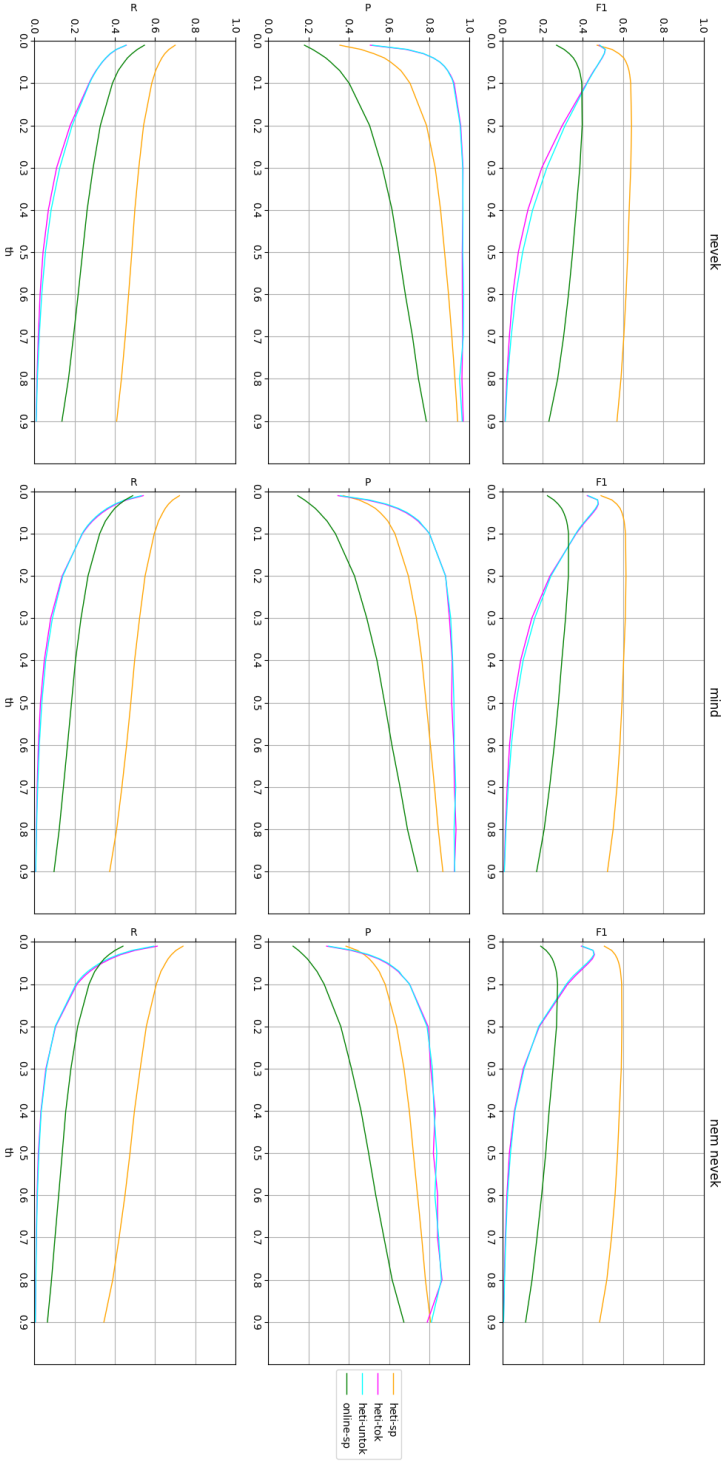
4. táblázat. A modellek betanítására és mérésére használt korpuszok jellemzői

Méréseink során a címkéket csoportokba osztottuk a tanítókorpuszbeli gyakoriságuk szerint. Mértük az egyes gyakorisági osztályokba tartozó címkék pontosságát, fedését és F_1-mértékét a javasolt címkelista különböző konfidenciaszintek melletti vágása esetén. Külön mértük a név-, illetve fogalmi címkékre, valamint az összes címkére vonatkozó eredő teljesítményt.

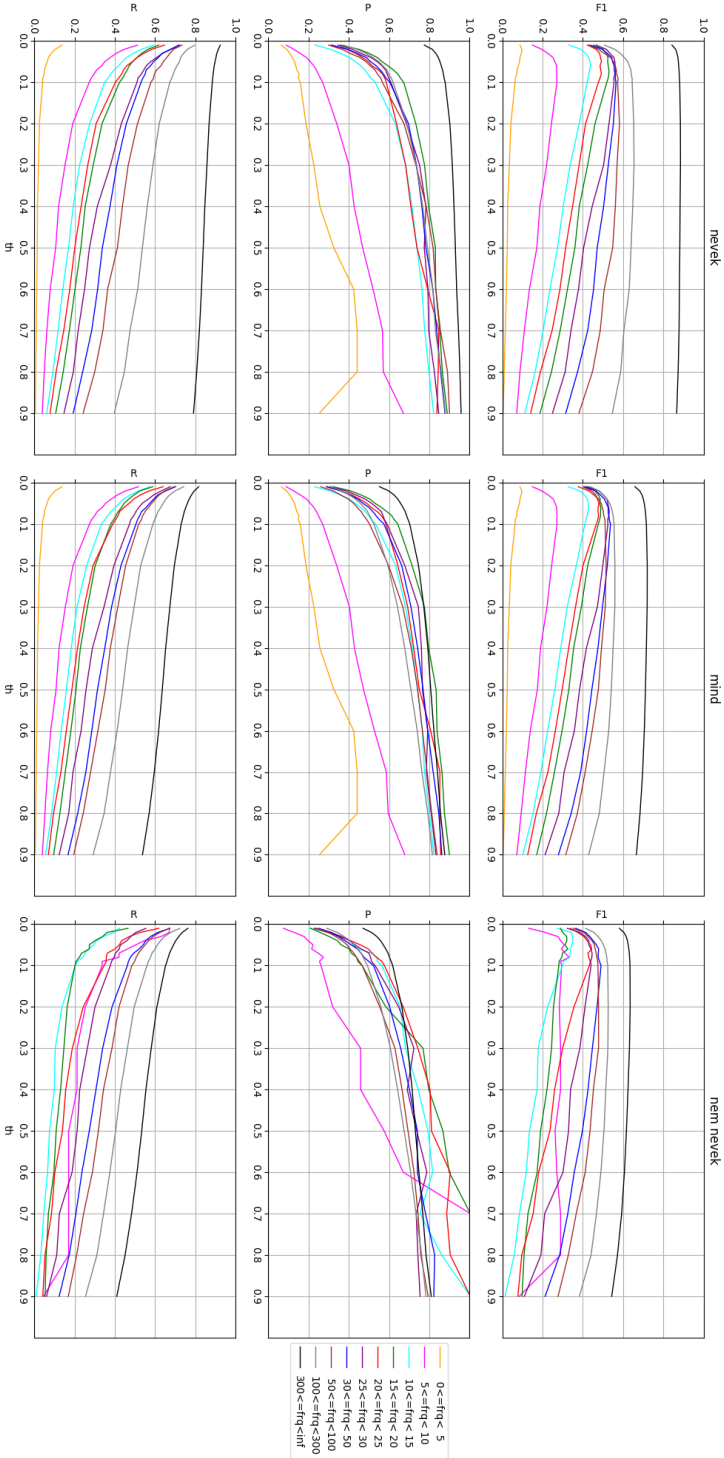
Méréseink eredményét a 2–5. ábrákon mutatjuk be. Az összes címkét figyelembe véve az egyes modellek pontosságának, fedésének és F-mértékének alakulását a vágási küszöb függvényében a 2. ábrán láthatjuk. Láthatóan a hetilapkorpuszon betanított SentencePiece tokenizálóval tokenizált *(heti-sp)* modell nyújtja a legjobb teljesítményt a fedés és az eredő F_1-mérték szerint. A hagyományos tokenizálóval tokenizált szövegen betanított modell *(heti-tok)* teljesítménye gyakorlatilag azonos a tokenizálatlan szövegen betanítottéval *(heti-untok)*. Ugyan ezek pontossága magasabb a *heti-sp* modellénél az alacsonyabb vágási konfidenciaértékeknél, fedésük viszont sokkal rosszabb. Az online korpuszon betanított modell *(online-sp)* mért teljesítménye az eredeti címkék sokkal nagyobb diverzitása miatt elmarad a *heti-sp* modellétől, de a címkék minőségével kapcsolatos szubjektív benyomás a megjelenő szinonim címkék miatt nem rosszabb, sőt a szerkesztőség munkatársai ezt modellt érezték jobbnak. Az online korpusz címkéinek normalizálására létrehoztunk egy eszközt, amelyet jelen kötet egy másik cikkében mutatunk be (Novák és Novák, 2020). A névcímkékre minden modell gyakorlatilag minden szempontból láthatóan jobb teljesítményt nyújt. Az egyetlen kivétel ez alól, hogy hetilapmodelleknél alacsony vágási értékeknél a fedés kicsit jobb a fogalmi címkékre, mint nevekre. F-mértékben a nem SP tokenizált modellek nagyjából a 0,02-es szintnél érik el a maximális teljesítményüket, az SP-modellek ezzel szemben 0,2-nél, de az egész tartományban viszonylag kiegyensúlyozott teljesítményt nyújtanak.

Megmértük a modellek eredő teljesítményét az egyes címkegyakorisági osztályokra külön-külön (3–5. ábrák) is. A *heti-sp* modell (3. ábra) az 5 alatti gyakoriságú nem névcímkék[4] kivételével mindegyik, az *online-sp* modell (4. ábra) pedig gyakorlatilag mindegyik gyakorisági osztályban felmutat valamilyen teljesítményt, bár a fedés a ritkább címkék körében nem túl magas. A 10 alatti gyakoriságú címkéknél a pontosság sem feltétlenül. Általában megfigyelhető, hogy minél gyakoribb egy címke, annál jobb a fedés és a pontosság is, de pl. a *heti-sp* modellnél az 5-9 gyakoriságú fogalmi címkék fedése relatíve kiemelkedik. A hagyományosan tokenizált *heti-tok* modellnél (5. ábra) csak elég alacsony vágási értékeknél kezd megjelenni mérhető fedés. Ennek köszönhetőek a pontossággal kapcsolatos kusza ábrák, hiszen itt a tesztanyagban előforduló viszonylag ritka (30 tanítóminta alatti gyakoriságú) címkék közül (ez a 7684 címkeelőfordulás teszi ki a tesztanyag címkeelőfordulásainak 1/5-ét) csak néhány (konkrétan összesen 40) jelenik meg a 0.05-ös vágási szint felett egyáltalán.

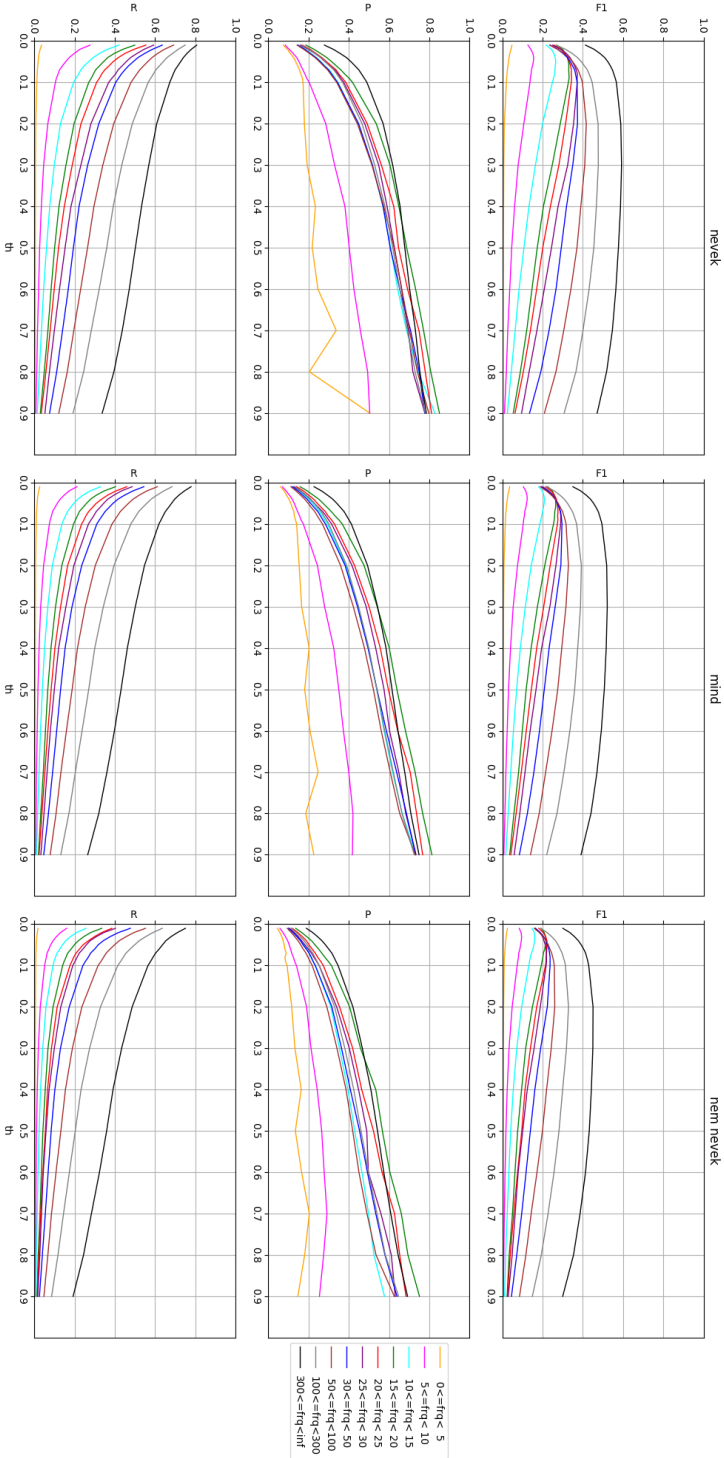
[4] Ilyen címkék nemigen vannak: a teljes tesztkorpuszban összesen 9 címkeelőfordulás tartozik ebbe az osztályba.



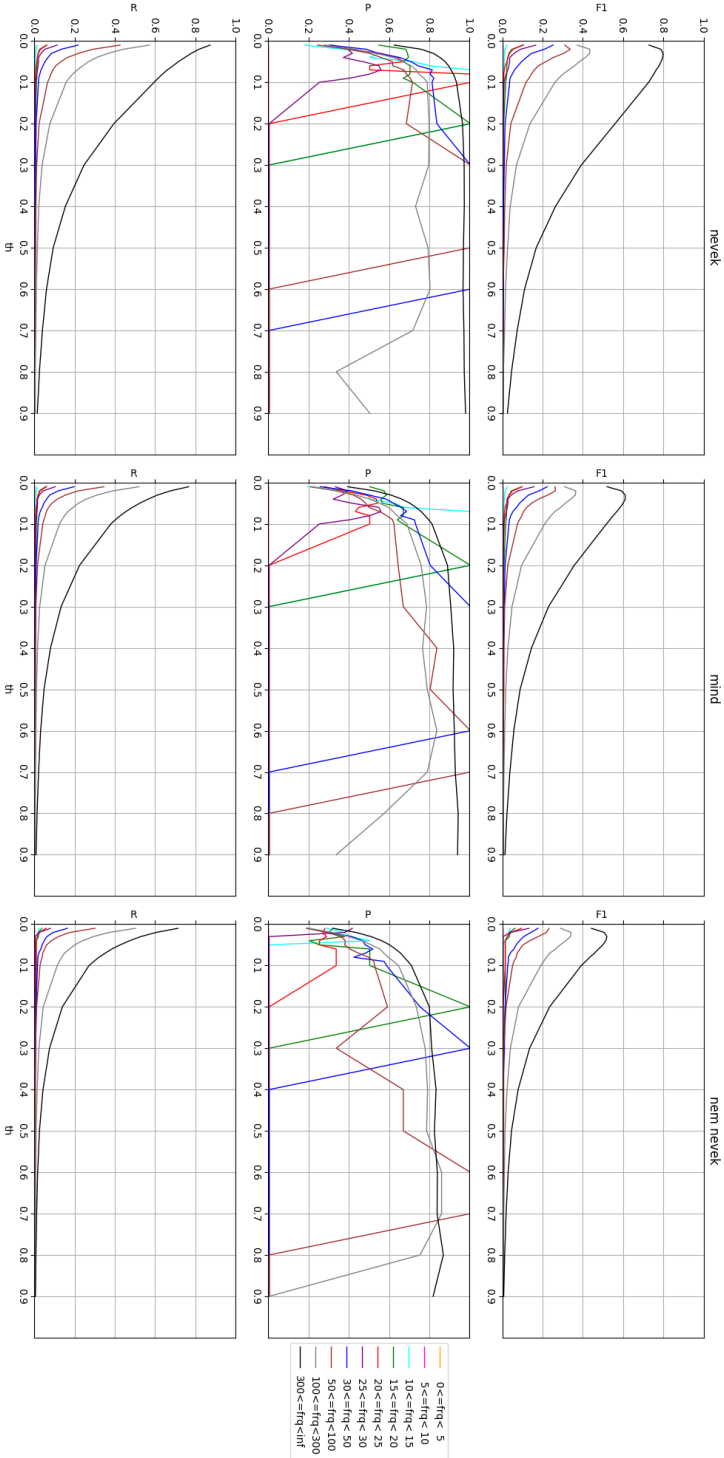
2. ábra: Négy modell teljes címkekészleten nyújtott teljesítménye a vágási küszöb függvényében: P, R, F_1 eredmények



3. ábra: Hetilapmodell, SentencePiece tokenizáló, címkegyakorisági osztályok szerinti P, R, F_1 eredmények a vágási küszöb függvényében



4. ábra: Online modell, SentencePiece tokenizáló, címkegyakorisági osztályok szerinti P, R, F_1 eredmények a vágási küszöb függvényében



5. ábra: Hetilapmodell, hagyományos tokenizálás, címkegyakorisági osztályok szerinti P, R, F_1 eredmények a vágási küszöb függvényében

7. Összegzés

Létrehoztunk egy címkézőrendszert, amellyel sajtószövegek automatikus tematikus címkézését tudjuk megvalósítani. A rendszerhez olyan felhasználói felületet hoztunk létre, amely lehetővé teszi a felhasználó számára a rendszer bizonyos paramétereinek (pl. az ajánlati lista vágását szabályozó konfidenciaszint) beállítását. Ennek segítségével az ajánlott kulcsszólista fedése és pontossága testre szabható. A rendszer segítségével legjobb esetben közel 80%-os pontossággal tudunk tematikus címkéket ajánlani sajtószövegek számára. A fastText osztályozót SentencePiece tokenizálóval kombinálva jelentősen tudtuk növelni a címkézőrendszer fedését, miközben a pontosság csökkenése tolerálható volt, ugyanakkor a modell mérete is töredékére csökkent. Illusztráltuk azt is, hogy a rendszer által ajánlott alacsonyabb konfidenciaértékű címkék, még ha nem szerepeltek is az eredeti címkék között, az esetek nagy részében jól illeszkednek a szöveg témájához.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 számú projekt keretében az FK 17 pályázati program valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Farkas, R.: Az origo automatikus címkézési projekt tapasztalatai. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009). pp. 84–92. Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2009)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. ACL, Valencia, Spain (2017)

Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 66–75. ACL, Melbourne, Australia (2018)

Novák, A., Novák, B.: Bu-bor-ék: grafikus címkenormalizáló eszköz. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2020)

Szántó, Zs., Vincze, V., Farkas, R.: Magyar nyelvű szó- és karakterszintű szóbeágyazások. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017). pp. 323–328. Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2017)

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. CoRR abs/1906.08237 (2019), http://arxiv.org/abs/1906.08237

Morfológia, helyesírás

The Role of Interpretable Patterns in Deep Learning for Morphology

Judit Ács[1,2], András Kornai[2]

[1] Department of Automation and Applied Informatics
Budapest University of Technology and Economics
[2] Institute for Computer Science and Control

Abstract. We examine the role of character patterns in three tasks: morphological analysis, lemmatization and copy. We use a modified version of the standard sequence-to-sequence model, where the encoder is a pattern matching network. Each pattern scores all possible N character long subwords (substrings) on the source side, and the highest scoring subword's score is used to initialize the decoder as well as the input to the attention mechanism. This method allows learning which subwords of the input are important for generating the output. By training the models on the same source but different target, we can compare what subwords are important for different tasks and how they relate to each other. We define a similarity metric, a generalized form of the Jaccard similarity, and assign a similarity score to each pair of the three tasks that work on the same source but may differ in target. We examine how these three tasks are related to each other in 12 languages. Our code is publicly available.[1]

1 Introduction

Deep neural networks are successful at various morphological tasks as exemplified in the yearly SIGMORPHON Shared Task (Cotterell et al., 2016, 2017, 2018). However these neural networks operate with continuous representations and weights which is in stark contrast with traditional, and hugely successful, rule-based morphology. There have been attempts to add rule-based and discrete elements to these models through various inductive biases (Aharoni and Goldberg, 2016).

In this paper we tackle two morphological tasks and the copy task as a control with an interpretable model, SoPa. Soft Patterns (Schwartz et al., 2018) or SoPa is a finite-state machine parameterized with a neural network, that learns linear patterns of predefined size. The patterns may contain epsilon transitions and self-loops but otherwise are linear. *Soft* refers to the fact that the patterns are intended to learn abstract representations that may have multiple surface representations, which SoPa can learn in an end-to-end fashion. We call these

[1] https://github.com/juditacs/deep-morphology

surface representations *subwords*, while the abstract patterns, *patterns* throughout the paper. An important upside of SoPa is that interpretable patterns can be extracted from each sample. (Schwartz et al., 2018) shows that SoPa is able to retrieve meaningful word-level patterns for sentiment analysis. Each pattern is matched against every possible subword and the highest scoring subword is recovered via a differentiable dynamic program, a variant of the forward algorithm. We apply this model as the encoder of a sequence-to-sequence or *seq2seq*[2] model (Sutskever et al., 2014), and add an LSTM (Hochreiter and Schmidhuber, 1997) decoder. We initialize the decoder's hidden state with the final scores of each SoPa pattern and we also apply Luong's attention (Luong et al., 2015) on the intermediate outputs generated by SoPa. We call this model SoPa Seq2seq. We compare each setup to a sequence-to-sequence with a bidirectional LSTM encoder, unidirectional LSTM decoder and Luong's attention.

We show that SoPa Seq2seq is often competitive with the LSTM baseline while also interpretable by design. SoPa Seq2seq is especially good at morphological analysis, often surpassing the LSTM baseline, which confirm our linguistic intuition namely that subword patterns are useful for extracting morphological information.

We also compare these models using a generalized form of Jaccard-similarity and we find that some trends coincide with linguistic intuition.

2 Data

Universal Morphology or UniMorph is project that aims to improve how NLP handles languages with complex morphology.[3] Specified in (Sylak-Glassman, 2016), UniMorph has been used to annotate 350 languages from the English edition of Wiktionary[4]. Wiktionary contains inflection tables that list inflected forms of a word. Part of the UniMorph project is converting these tables into *(lemma, inflected form, tags)* triplets such as *(ablak, ablakban, N IN+ESS SG)*. The first tag is the part-of-speech which is limited to the main open classes (nouns, verbs and adjectives) in most languages, `IN+ESS` is the inessive case and `SG` denotes singular.

2.1 Data sampling

Our goal is to sample 10000 training, 2000 development and 2000 test examples. We retrieved 109 UniMorph repositories (109 languages) but only 57 languages have at least 14000 samples, the lowest possible number for our purposes. We first prefilter the languages and assign them to languages families and genus using the World Atlas of Languages or WALS[5]. WALS does not contain extinct, constructed or liturgical languages, and we do not incorporate these in

[2] also called encoder-decoder model
[3] https://unimorph.github.io/
[4] https://en.wiktionary.org/
[5] https://wals.info/

Language	Family	Genus	sample	lemma	paradigm	alphabet	F/L	POS
Arabic	Afro-Asiatic	Semitic	138k	4007	196	45	26.3	NVA
Turkish	Altaic	Turkic	213k	3017	186	46	54.7	NVA
Quechua	Hokan	Yuman	178k	1003	553	22	146.8	NVA
Albanian	Indo-European	Albanian	14k	587	59	27	17.4	NV
Armenian	Indo-European	Armenian	259k	6991	134	46	35.3	NVA
Latvian	Indo-European	Baltic	129k	7238	78	34	10.3	NVA
Lithuanian	Indo-European	Baltic	33k	1391	139	56	20.1	NVA
Irish	Indo-European	Celtic	45k	7299	53	53	3.3	NVA
Danish	Indo-European	Germanic	25k	3190	14	44	7.7	NV
German	Indo-European	Germanic	171k	15032	37	63	4.5	NV
English	Indo-European	Germanic	115k	22765	5	65	4.0	V
Icelandic	Indo-European	Germanic	76k	4774	44	54	10.9	NV
Greek	Indo-European	Greek	147k	11872	118	76	6.5	NVA
Kurdish	Indo-European	Iranian	203k	14143	128	61	14.3	NVA
Asturian	Indo-European	Romance	29k	436	223	32	49.5	NVA
Catalan	Indo-European	Romance	81k	1547	53	35	40.6	V
French	Indo-European	Romance	358k	7528	48	44	35.3	V
Bulgarian	Indo-European	Slavic	54k	2413	95	31	18.9	NVA
Czech	Indo-European	Slavic	109k	5113	147	62	10.0	NVA
Slovenian	Indo-European	Slavic	59k	2533	94	56	8.9	NVA
Georgian	Kartvelian	Kartvelian	74k	3777	109	33	17.5	NVA
Adyghe	NW Caucasian	NW Caucasian	20k	1635	30	40	11.9	NA
Zulu	Niger-Congo	Bantoid	49k	566	249	46	57.2	NVA
Khaling	Sino-Tibetan	Mahakiranti	156k	591	432	32	91.5	V
Estonian	Uralic	Finnic	27k	886	64	26	28.0	NV
Finnish	Uralic	Finnic	1M	57165	97	50	27.1	NVA
Livvi	Uralic	Finnic	63k	15295	104	55	4.0	NVA
Northern Sami	Uralic	Saami	62k	2103	80	31	25.9	NVA
Hungarian	Uralic	Ugric	517k	14883	93	53	34.1	NV

Table 1. Dataset statistics. The languages are sorted by language family. F/L refers to the form-per-lemma ratio. POS indicates which part of speech are present in the dataset out of the nouns, verbs and adjectives.

our dataset. Out of the 109 languages, 19 have no WALS entry. 29 languages have large enough UniMorph datasets that allow obtaining 10000/2000/2000 samples.[6] Table 1 summarizes the dataset.

3 Tasks

We train both kinds of seq2seq models on three tasks: morphological analysis (abbreviated as *morphological analysis*), *lemmatization*, and *copy* or autoencoder. The source sequence is the inflected form of the word in all three tasks, while the target sequence is a list of morphosyntactic tags for morphological analysis, the lemma for lemmatization and the same as the source side for copy. Table 2 shows examples for the three tasks.

Inflected words and lemmas are treated as a sequence of characters but tags are treated as standalone symbols. We share the vocabulary and the embedding between the source and target side when training for copy and lemmatization but we use separate vocabularies for morphological analysis.

[6] Albanian has only 1982 test samples but we wanted to include it as a language isolate from the Indo-European family.

Language	Task	Source	Target
Hungarian	morphological analysis	vásároljanak	V SBJV PRS INDF 3 PL
Hungarian	morphological analysis	lepkékben	N IN+ESS PL
English	morphological analysis	hugging	V V.PTCP PRS
French	morphological analysis	désinstalleriez	V COND 2 PL
Hungarian	lemmatization	vásároljanak	vásárol
Hungarian	lemmatization	lepkékben	lepke
English	lemmatization	hugging	hug
French	lemmatization	désinstalleriez	désinstaller
Hungarian	copy	vásároljanak	vásároljanak
Hungarian	copy	lepkékben	lepkékben
English	copy	hugging	hugging
French	copy	désinstalleriez	désinstalleriez

Table 2. Dataset examples.

4 Models

We train two kinds of sequence-to-sequence models which only differ in the choice of the encoder. Both models first pass the input through an embedding. We train the embeddings from randomly initialized values and do not use pretrained embeddings. We use character embeddings with 50 dimensions for character inputs and outputs and tag embeddings with 20 dimensions for morphological tags (only for morphological analysis). The embeddings are shared between the encoder and decoder for lemmatization and copy, since both the source and the target sequences are characters. The output of the source embedding is the input to the encoder module which is a SoPa with 120 patterns in SoPa Seq2seq case and a bidirectional LSTM in the baseline. The decoder later attends on the intermediary outputs of these modules. The final hidden state of the encoder module is used to initialize the decoder. The decoder side of these models is identical in both setups, an LSTM with Luong's attention. All LSTMs have 64 hidden cells and a single layer.

The size of SoPa patterns (3, 4, and 5 in our case) define the number of forward arcs that a pattern has. These may contain epsilon steps and self loops but an epsilon or a self loops is always followed by a main transition (consuming an actual symbol). This means that a 3 long pattern may contain one epsilon and one main transition, two epsilons or two main transitions. Any main transition may be preceded by a self loop. The pattern size includes the start state and the end state. In our experiments we used 3, 4, and 5 long patterns, 40 patterns of each length.

Most of the training details are also identical. We train with batch size 64, and we use early stopping if the development loss and accuracy stop improving for 5 epochs. We maximize the number of epochs in 200 but this is never reached. We save the best model based on development accuracy. We use the Adam optimizer with 0.001 learning rate for all experiments.

SoPa is more difficult to train than LSTMs, so we decay the learning rate by 0.5 if the development loss does not decrease for 4 epochs.

5 Model similarity

We define a similarity metric between two SoPa Seq2seq models measured on datasets that share their source side. The target side may differ. The three tasks introduced in Section 3, all take inflected word forms as their source sequence, which allows computing our similarity metric between each pair of tasks.

SoPa works with a predefined number of patterns and tries matching each pattern on any subword of the input with a particular length. The highest scoring subword is used in the final source representation. We take the highest scoring $T = 10$ patterns for each input and compare the subwords that resulted in these scores. The metric is defined as the average similarity over the dataset D:

$$\mathrm{Sim}(M_1, M_2, D) = \frac{1}{|D|} \sum_{d \in D} S(M_1(d), M_2(d)), \tag{1}$$

where M_1 and M_2 are the models, and S is the similarity of the two representations generated by the encoder side of the models on sample d, defined as:

$$S(M_1(d), M_2(d)) = \frac{1}{2T}\Big(\sum_{p_i \in P_1} \max_{p_j \in P_2} J(p_i, p_j) + \sum_{p_j \in P_2} \max_{p_i \in P_1} J(p_i, p_j)\Big), \tag{2}$$

where T is a predefined number of highest scoring patterns on that sample (10 in our experiments), P_1 is the set of T highest scoring patterns of M_1, P_2 is the set of T highest scoring patterns of M_2 and J is the Jaccard similarity of two subwords defined as the proportion of overlapping symbols by the union of all symbols. Jaccard similarity is 0 if there is no overlap and is 1 when the subwords are the same. For each sample, we first choose the highest scoring T patterns from each model, we denote these sets of patterns as P_1 and P_2. Then we find the subwords corresponding to these patterns. We compute the pairwise Jaccard similarities between every element of P_1 and P_2. Then for each pattern, we find the most similar pattern from the other model. The average of these scores is the similarity of the two models on that sample (see Eq. 2) and the average over all samples (see Eq. 1) is the similarity of two models on dataset D. This metric is symmetric and it ranges from 0 to 1. Table 3 shows a small example of this similarity on the word *ablakban*.

6 Results and analysis

We first show that SoPa Seq2seq is competitive with the LSTM Seq2seq baseline, especially for morphological analysis. An output is considered accurate if it fully matches the reference and we do not consider partial matching. Some

	ˆablakban$	ˆablakban$	ˆablakban$	ˆablakban	Max
ˆablakban$	0	0.2	1	0.75	1
ˆablakban$	0	0.5	0.5	0.75	0.75
ˆablakban$	0	0.5	0	0.167	0.5
ˆablakban$	0	0.75	0.167	0.333	0.75
Max	0	0.75	1	0.75	J=0.6875

Table 3. Simlarity (Eq. 2) between two models M_1 and M_2 on one sample using the 4 highest scoring subwords ($T = 4$) with the subwords underlined. Rows correspond to the highest scoring subwords from M_1 (ban, kba, lak, kban), while columns correspond to the subwords from M_2 (ˆab, akb, ban, lakb). A Jaccard similarity matrix (with position information) is constructed. The final similarity is the mean maximum of every row and every column of the matrix.

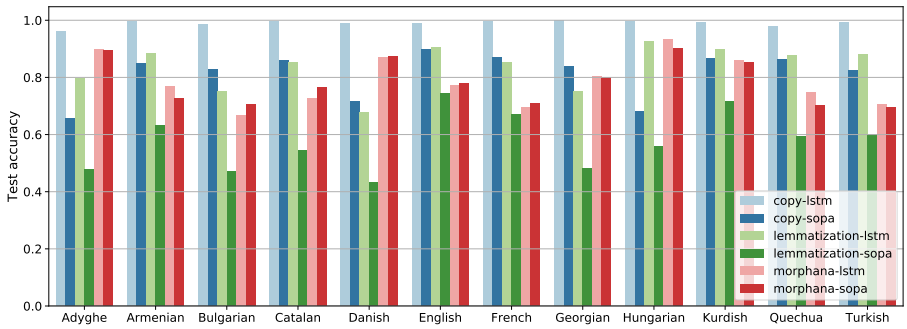


Fig. 1. Accuracy of SoPa Seq2seq models on each language and task.

languages prove to be too difficult for the models, which may be due to the lack of context that is often needed for morphological analysis and orthographic changes often present in lemmatization. We continue our analysis on languages where each of the three tasks are performed by SoPa 'reasonably well', which we set to 40% accuracy or higher on the development set. This leaves us with 12 languages. The reason we set a lower limit to accuracy is that we have no reason to believe that a bad model's representation is useful for the task. Fig. 1 shows the test accuracy in these languages. Lemmatization is consistently the most difficult task for SoPa, while SoPa is on pair with LSTM Seq2seq in morphological analysis, sometimes outperforming it. We attribute this result to the fact that a morphological tag often corresponds to a single morpheme, usually with a few possible surface realizations that SoPa's 'soft' patterns can pick up on. On the other hand lemmatization and copy require regenerating much of the input which is more difficult from an inherently summarized representation such as the one SoPa generates.

We continue by computing the pairwise similarity value defined in Eq. 1 between the three tasks. Higher values indicate that SoPa finds similar patterns valuable for generating the output. Fig. 2 shows the pairwise similarity of models

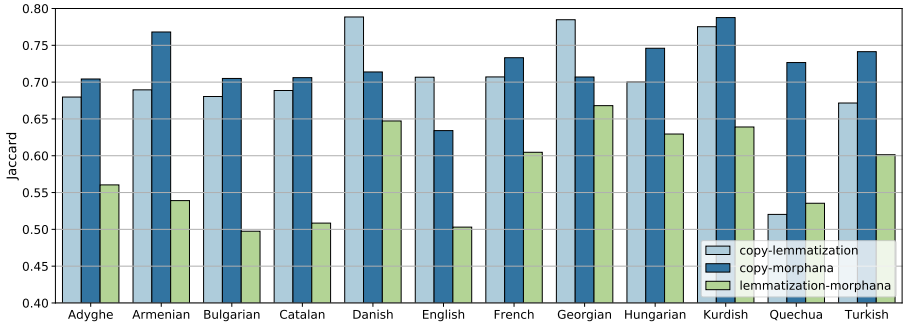


Fig. 2. Model similarity between all task pairs by language. Higher similarity indicates that two models handle the same source in a more similar way.

trained for the three tasks. We only compute these similarities on samples where the output of *both* models are correct (generally 40-60% of the test samples).

Lemmatization and morphological analysis are the least similar in almost every language. This is not surprising considering that lemmatization is the task of discarding information that morphological analysis needs to correctly tag. Quechua is the only exception from this trend which could be explained by the very rich inflectional morphology (especially at the type-level) that results in lemmas being significantly shorter than inflected forms. This means that copy needs to memorize a lot more of the source word than lemmatization.

Another trend we observe, is that copy and morphological analysis are more similar than copy and lemmatization in languages with rich inflectional morphology such as Armenian, Hungarian, Kurdish and Turkish and the opposite is true in fusional and morphologically poor languages such as Danish and English. Georgian seems to be an exception and further exploration is an exciting research direction.

Finally we demonstrate SoPa's interpretability by extracting the most frequently matched subwords in each language and task. Table 4 lists the most common subwords in English, French and Hungarian in each task. It should be noted that these subwords are very short because we used 3, 4 and 5 long patterns that match 2, 3 and 4 characters not including self loops and short patterns simply occur more frequently.

7 Conclusion

We presented an application of Soft Patterns – a finite state automaton parameterized by a neural network – as the encoder of a sequence-to-sequence model. We show that it is competitive with the popular LSTM encoder on character-level copy and morphological tagging, while providing interpretable patterns.

We analyzed the behavior of SoPa encoders on morphological analysis, lemmatization and copy by computing the average Jaccard similarity between

language	task	subwords
English	copy	ed,e$,ed$,es,in,at,re,s$,te,ri
English	lemmatization	at,g$,er,in,ng,iz,s$,en,ize,es
English	morphological_analysis	d$,s$,e$,es$,$,ed,ed$,o,ng,g$
French	copy	s$,ss,is,as,ie,ai,z$,nt,ns,en
French	lemmatization	er,s$,t$,nt,ie,ns,ra,is,ri,^d
French	morphological_analysis	s$,t$,z$,nt$,ez$,e$,ai,er,ns$,es$
Hungarian	copy	l$,n$,k$,sz,t$,nk$,kk,el,ok,na
Hungarian	lemmatization	sz,t$,k$,l$,ta,tá,^k,n$,kb,ró
Hungarian	morphological_analysis	l$,t$,n$,k$,ek,a$,$,g$,á$,ak$

Table 4. Top subwords extracted from English, French and Hungarian. ^and $ denote word start and end respectively.

the patterns extracted from the source side. We found two trends that coincide with linguistic intuition. One is that lemmatization and morphological analysis require patterns that match less similar subwords than the other two task pairs. The other one is that copy and morphological analysis are more similar in languages with rich inflectional morphology.

Acknowledgments

Work partially supported by 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence; and National Research, Development and Innovation Office grant NKFIH #120145 'Deep Learning of Morphological Structure'. We thank Roy Schwartz for his help in understanding the inner mechanics of SoPa.

Bibliography

Aharoni, R., Goldberg, Y.: Morphological inflection generation with hard monotonic attention. arXiv preprint arXiv:1611.01487 (2016)

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A.D., Kann, K., Mielke, S., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J., Hulden, M.: The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In: Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. Association for Computational Linguistics, Brussels, Belgium (October 2018)

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., Hulden, M.: The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In: Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection. Association for Computational Linguistics, Vancouver, Canada (August 2017)

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., Hulden, M.: The SIGMORPHON 2016 shared task—morphological reinflection. In: Proceedings of the 2016 Meeting of SIGMORPHON. Association for Computational Linguistics, Berlin, Germany (August 2016)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (Nov 1997)

Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421. Association for Computational Linguistics (2015), http://www.aclweb.org/anthology/D15-1166

Schwartz, R., Thomson, S., Smith, N.A.: SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines. In: Proc. 56th ACL Annual Meeting. pp. 295–305. Melbourne, Australia (2018)

Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proc. NIPS. pp. 3104–3112. Montreal, CA (2014), http://arxiv.org/abs/1409.3215

Sylak-Glassman, J.: The composition and use of the universal morphological feature schema (unimorph schema). Tech. rep. (2016)

Automatikus ékezetvisszaállítás transzformer modellen alapuló neurális gépi fordítással

Laki László János[1,2], Yang Zijian Győző[1,2]

[1]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/a.
{laki.laszlo, yang.zijian.gyozo}@itk.ppke.hu

Kivonat Cikkünkben egy ékezetvisszaállító programot mutatunk be, amelyet a jelenkori „state-of-the-art" transzformer modellen alapuló neurális gépi fordító rendszerrel tanítottunk be. A mobil eszközökön történő üzenetírás elterjedésével és a minél gyorsabb szövegbevitelre való törekvéssel tömeges jelenséggé vált az ékezetes betűk elhagyása a gépelt írásban. Ennek egyik következménye, hogy a interneten elérhető – főleg a szociális médiából származó – korpuszok egy része ékezetmentes. Egy ékezetvisszaállító program segítségével vissza tudjuk állítani az ékezethiányos szavakat, valamint integrálva szövegbeviteli eszközökkel támogatni tudjuk a felhasználók számára a szövegbevitelt. Az általunk létrehozott rendszer, annak ellenére, hogy semmilyen morfológiai elemzőt nem használ, több mint 99,7%-os pontossággal tudja helyesen visszaállítani az ékezeteket magyar nyelv esetében. A hibaanalízis során kiderült, hogy a hibák több mint 50%-a a többértelműségből fakad, illetve, hogy a rendszerünk által ajánlott ékezetesítés utáni mondat is helyes. Készítettünk egy demó felületet is, amelyen ki lehet próbálni a különböző modellek működését.
Kulcsszavak: ékezetvisszaállítás, neruálisháló-alapú gépi fordítás, NMT, transzformer modell

1. Bevezetés

Napjaink számítógépes nyelvészei számára nagy lehetőséget nyújtanak az interneten elérhető nagy mennyiségű szövegek. Számos részterületen használjuk a weboldalakról összegyűjtött korpuszokat, mint például a gépi fordítás, a szöveg kivonatolás vagy az érzelem detektálás. Ezekhez a feladatokhoz viszont nélkülözhetetlen, hogy a vizsgált szöveg a lehető legjobb minőségű legyen.

A mobil eszközökön írt szövegek és üzenetek esetében tömegjelenséggé vált az ékezetes betűk elhagyása. Ennek következményeképp léteznek olyan korpuszok is, amelyek egy része ékezetmentes, így nem működnek rajtuk a természetes szövegen betanított szövegfeldolgozó modellek. Egy ékezetvisszaállító program segítségével vissza tudjuk állítani az ékezethiányos szavakat, valamint integrálva további szövegbeviteli eszközökkel támogatni tudjuk a felhasználók szövegbevitelét. A feladat komplexitását a többértelmű szavak visszaállítása is növeli.

Az elmúlt években a neurálishálózat-alapú módszerek eredményei túlszárnyalták az addigi legjobb rendszereket. Ez a nyelvtechnológia területén is megmutatkozik, ezért célunk az volt, hogy megvizsgáljuk az ékezetesítés problémáját a jelenlegi „state-of-the-art" NMT-alapú rendszerrel.

2. Kapcsolódó munkák

Az elmúlt években több kísérlet is született az ékezetek helyreállítását megcélozva. Nyelvfüggetlen módszerekkel kísérletezett Mihalcea és Nastase (Mihalcea és Nastase, 2002). Kutatásukban gépi tanulásos módszereket alkalmaztak a probléma megoldására. Az egyik módszer, amikor az ékezetes betűk pozíciója és környezete segíti a megoldást. Ezzel a megközelítéssel 95%-os pontosságot értek el. Egy másik módszerükkel korpuszból becsülték meg a különböző ékezetes szavak disztribúcióját, mellyel 98%-os pontosságot értek el. A rendszer hátránya viszont, hogy a korpuszban nem szereplő – ismeretlen szavakat – nem tudja kezelni.

Charlifter szintén nyelvfüggetlen megoldást keresett (Scannell, 2011). Lexikonalapú statisztikai módszerekkel állítja helyre az ékezetet. Figyeli a közvetlen környezetet és az ismeretlen szavak kezelésére karakteralapú statisztikai modellt alkalmaz. A legjobb esetben is csak 93%-os pontosságot ért el.

Nyelvspecifikus kutatásokat végzett Yarowsky spanyol és francia nyelvre (Yarowsky, 1999), valamint Zweigenbaum és Grabar francia nyelvre (Zweigenbaum és Grabar, 2002).

Magyar nyelvre Németh és társai (Németh és mtsai, 2000) egy text-to-speech alkalmazást mutatnak be, melyben kezelik az ékezethiányos szavakat. A probléma megoldásához morfológiai és szintaktikai elemzőt is használnak, mellyel 95%-os pontosságot értek el. Novák és Siklósi (Novák és Siklósi, 2015, 2016) statisztikai gépi fordítást (SMT) alkalmaznak az ékezet helyreállításához. Morfológiai elemző nélküli és egy morfológiai elemzővel rendelkező SMT-vel is végeztek kísérleteket. A legjobb eredményt – 99,06% – a morfológiai elemzővel érték el.

Nagy Péter szakdolgozatában (Nagy, 2018) RNN-alapú neurális gépi fordítórendszert alkalmazott a feladat megoldására. Legjobb eredménye eléri a 99,5%-os pontosságot. Munkája során BPE (Byte pair encoding) tokenizálást (Sennrich és mtsai, 2015) végzett úgy, hogy külön modellt használt a forrás- és a célnyelvi tanítóanyagra. Ez a dolgozat tekinthető a leghasonlóbbnak saját munkánkhoz. Kutatásunkban az RNN modell helyett a jelenlegi „state-of-the-art" transzformer modellt használjuk, míg a BPE helyett a közös szótárral rendelkező Sentence Piece (SPM) tokenizálást alkalmazunk. Ezzel a technikával sikerül további javulást elérnünk.

3. Ékezetes szavak helyreállítása

A korpusz-alapú gépi fordító rendszer lényege, hogy transzformációt képez tetszőleges forrás- és célnyelvi mondatok között, ahol a rendszer betanításához nem

kell más, mint egy kétnyelvű párhuzamos korpusz. Az ékezetes szavak helyreállítására kézenfekvő választás a gépi fordítás módszereit használni, mivel az ékezetlen és az ékezetes mondatok grammatikailag, szókincsileg és szószerkezetileg nagyon hasonlóak.

A neurális hálózat tanításához nagy mennyiségű tanítóanyagra van szükség, melynek előállítása a jelen feladathoz igen könnyű. A tanítóanyag létrehozásához annyit kellett tenni, hogy egy egynyelvű korpusz ékezetes karaktereiről szabály alapon eltávolítottuk az ékezeteket.

3.1. A korpusz

A korpusz létrehozásához az online elérhető Open Subtitles[1] nevű angol-magyar párhuzamos korpuszának magyar oldali szövegét használtuk. A korpusz TV és mozi filmekre létrehozott feliratokból áll. Ennek megfelelően főleg rövidebb, informális mondatokat tartalmaz. A gépi fordító rendszer célnyelvi korpuszának előállításához a mondatokban az ékezetes karaktereket lecseréltük az ékezet nélküli párjukra (pl. á→a, é→e stb.)

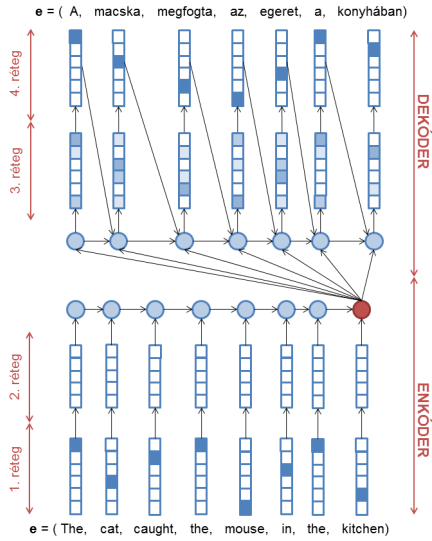
A korpusz megközelítőleg 29 millió szegmensből áll, melyből 5000 mondatot validációs és 3000 mondatot tesztelési célra elkülönítettünk. A korpusz az egyik legnagyobb szabadon hozzáférhető párhuzamos tanítóanyagnak számít, ellenben mérete elmarad az egynyelvű tanítóanyagokétól. Választásunk azért esett erre az adathalmazra, mert több párhuzamos kutatásunk során is használjuk, és néhány koprusztisztító lépést már előzetesen eszközöltünk rajta. Kivettük azokat a mondatokat, amelyek speciális karaktereket (pl. kínai, japán, cirill stb.) tartalmaztak, valamint a teszt halmaz mondatait kézzel kijavítottuk. Mérete elégséges a neurális hálózatok helyes betanítására, valamint a tanítási idő is viszonylag kezelhető marad (1-2 nap).

3.2. A neurális gépi fordítórendszer

A 2010-es évek első felére a statisztikai gépi fordítórendszerek elérték teljesítőképességük határát. Az alapjait képező módszert és a létrehozott keretrendszereket a kutatók nagyon sok befektetett munka ellenére lényegében nem sikerült tovább javítani. Az áttörést (Bahdanau és mtsai, 2015) rendszere hozta el, ami egy figyelmi modellel támogatott enkóder-dekóder architektúrájú NMT rendszer volt. A modell lényege, hogy kettéválasztja a fordítás folyamatát két elkülöníthető részre. A kódolás során lényegében egy RNN-alapú seq2seq modellt hoz létre, tehát a szóbeágyazási modellhez hasonlóan a fordítandó modellekből egy n-dimenziós vektort készít. Az 1. ábrán ez a vektor felel meg az ábra közepén látható piros/sötét node-nak. A második fázis a dekódolás, ahol a mondatvektorból generálja ki a célnyelvi mondatot egy RNN réteg segítségével.

Innentől számítva az NMT rendszerek átvették a vezető szerepet az SMT-től. 2017-ben a Google cég munkatársai (Vaswani és mtsai, 2017) publikálták és szabadon hozzáférhetővé tették az úgynevezett multi-attention réteggel támogatott

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php



1. ábra: Enkóder-dekóder architektúra vázlatos rajza

NMT rendszerüket. Ezt a szakirodalomban transzformer-alapú architektúrának nevezik. A módszer lényege, hogy az eddigi egy helyett több figyelmi réteget helyeztek el a rendszerben, ami segítségével nagymértékben nőtt a többértelmű szavak fordításának minősége.

Munkánk során a Marian NMT(Junczys-Dowmunt és mtsai, 2018) nevű keretrendszert használtuk, ami egy c++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően [2] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

3.3. A Sentence Piece tokenizáló

Az NMT rendszerek működése GPU processzorokon történik, melyek egyik szűk keresztmetszete a bennük található memória mérete. Ez határozza meg a létrehozható NMT rendszer szótárának a méretét. Egy szóalapú rendszer esetében az általánosságban 100K különálló szóban korlátozzák le a rendszert, így a további szavakat ismeretlenként kezeli.

(Sennrich és mtsai, 2015) ezt a problémát úgy oldották meg, hogy a szavak helyett úgynevezett subword (szótöredék) szintre csökkentették a legkisebb fordítási egységet. A BPE (Byte Pair Encoding) egy adattömörítő eljárás, ahol a leggyakoribb bájtpárokat egy olyan bájttal helyettesítjük, amely nem szerepel

[2] https://marian-nmt.github.io/

magában az adatban. Az eljárás a korpuszon először egy karakteralapú szótárat hoz létre, ahol minden szót karakterek sorozataként ábrázol. Ezután gyakoriság alapján a gyakori karaktersorozatokat önálló tokenekként kezeli. Ezzel az adat tömörítése mellett az ismeretlen szavak kezelését is megoldja, hiszen a részszavakból előállítható egy olyan összetétel, amely nem szerepelt eredetileg a korpuszban.

Ezt a módszert fejlesztették tovább (Kudo és Richardson, 2018). Az általuk létrehozott Sentence Piece nevű eszköz egy felügyelet nélküli szöveg tokenizáló és detokenizáló, melyet elsősorban a neurálishálózat-alapú gépitanulásos feladatokhoz fejlesztettek ki. Implementálva van benne a BPE metrika, ami egy unigram nyelvmodellel (Kudo, 2018) van súlyozva. Használatával elhagyhatók a költséges nyelvspecifikus előfeldolgozási lépések, mint például a tokenizálás vagy a kisbetűsítés. A módszer lényege, hogy a természetes szöveget úgy alakítja át, hogy abban a különböző „szavak" száma korlátos legyen, valamint az így létrejött tanítóanyagban nem lesznek ismeretlen szavak. Ennek köszönhetően a neurális hálózatok paraméterszáma nagymértékben csökkenthető.

(1) Sima szöveg: Petőfi Sándor egy nagyszerű költő.
SPM szöveg: Pető fi □ S ándor □egy □nagyszerű □költő .

A fenti példában látható az SPM modell kimenete. A sima szöveg szavait gyakran előforduló karakter sorozatokra tördeli szét. Érdekes megfigyelni, hogy az eredeti mondat szóközeit is a szavakhoz csatolja és mint önálló karaktert (□) kezeli.

3.4. Megjelenítő felület

Készítettünk egy demó felületet[3], amelyen ki lehet próbálni a különböző modellek működését. Egy lenyíló menün keresztül lehet kiválasztani a tesztelendő modellt, majd egy input mezőn begépelhetjük a szavakat. A szóközök után megvizsgálja az addig leírt szövegrészletet, és ha hibásnak találja, ajánlatot tesz a javításra.

4. Kísérletek

A minőségbeli összehasonlíthatóság végett betanítottuk (Novák és Siklósi, 2015) által leírt morfológiai elemző nélküli SMT-t, valamint (Nagy, 2018) által jegyzett RNN-alapú neurális gépifordító rendszert (NMT-RNN) is.

Az SMT tanításához a Moses nevű keretrendszert (Koehn és mtsai, 2007) használtuk, ahol a nyelvmodellt a KenLM-el (Heafield, 2011) hoztuk létre. A rendszer tanítása során az alapbeállításokat használtuk és kihagytuk a szóösszerendelő és az átrendező lépéseket. Ezekre a lépésekre nem volt szükségünk, hiszen azonos a szószám és a szósorrend monoton a forrás- és a célnyelvi oldalon.

[3] http://nlpg.itk.ppke.hu/projects/accent

A fordítás előfeldolgozó fázisában történik egy tokenizáció és egy „truecase" lépés. A truecase-ing egyfajta kisbetűsítés, ahol a mondat kezdő szaváról döntjük el, hogy azt alapesetben kis- vagy nagybetűs formában használjuk. A fordítás utófeldolgozása során történik egy „detruecase" lépés és egy detokenizáció.

Az NMT tanításához a Marian neurális gépi fordítórendszert használtuk. Az RNN-alapú NMT beállításhoz a (Nagy, 2018) dolgozatában szereplő értékeket vettük alapul. A rendszer fontos jellemzője, hogy a BPE modellt külön tanítja be a forrás- és a célnyelvi korpuszokból. Ennek az a jelentősége, hogy a két esetben ugyanazt a szót a forrás- és a célnyelvi oldalon különböző subword formában tördelheti a rendszer. Emiatt sérül a szavak egy-egyértelmű megfeleltethetősége. A rendszer másik tulajdonsága, hogy 5 egymást követő sikertelen iteráció után megáll (early-stopping).

Munkánk elméleti alapját az újonnan elérhető transformer és SPM technológiák adták. Kíváncsiak voltunk arra, hogy ékezetesítés esetén is sikerül-e elérni a rendszerek természetes nyelveken bemutatott minőségi javulását. A továbbiakban NMT-TM néven hivatkozunk a saját modellünkre. A rendszerünk tanításához az alábbi paramétereket használtuk:

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0,1;
- learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
- optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
- label-smoothing: 0,1; exponential-smoothing

5. Eredmények

Kutatásunk során megmértük a gép által adott szóalapú eredmény pontosságát (precision), fedését (recall) és az abszolút pontosságot (accuracy). Mivel a gépi fordítás során az eredetileg helyes szavak is megváltozhatnak, szükséges megvizsgálni a fordítás pontosságát az összes szóra (ALL). Ezenkívül elvégeztük a kiértékelést azokra a szavakra nézve is, amelyek rendelkeznek magánhangzóval (MGH), vagyis a feladatra nézve releváns szavakat.

	ALL			MGH		
	Pontosság (Precision)	Fedés (Recall)	Abs pontosság (Accuracy)	Pontosság (Precision)	Fedés (Recall)	Abs pontosság (Accuracy)
SMT	97,96%	96,97%	98,49%	98,13%	97,04%	98,49%
NMT-RNN	97,04%	97,54%	98,58%	97,16%	97,60%	98,56%
NMT-TM	**99,38%**	**99,28%**	**99,63%**	**99,42%**	**99,33%**	**99,62%**

1. táblázat. A különböző ékezetesítő modellek kiértékelése

A 1. táblázat eredményei alapján láthatjuk, hogy az általunk létrehozott rendszer, amely transzformer modellt és Sentence Piece tokenizálót használ, min-

den esetben a legjobb eredményt ért el. Érdemes megemlíteni, hogy annak ellenére, hogy pontosságban (precision) az SMT jobb eredményt ért el az NMT-RNN modellhez képest, fedésben (recall) és a rendszer pontosságát (accuracy) illetően az NMT-RNN teljesített jobban.

Az NMT-TM modell teljesítménye minden esetben meghaladja a 99,27%-ot. Összesített rendszerszintű pontossága eléri a 99,63%-ot.

6. Hibaelemzés

Az eredmények mélyebb elemzése során megvizsgáltuk a rendszer által elkövetett hibákat. A 2. táblázatban láthatjuk a hibatípusokat. A tesztanyag 3000 mondatából (18438 token és 8957 type) mindössze 67 mondatban volt hiba, melyekben összesen 69 darab szót vélt hibásnak. Ezen hibák mélyebb elemzéséből láthatjuk, hogy nagy részük nem tekinthető valódi hibának. Az egyik ilyen hibakör a szövegkörnyezet ismerete nélküli többértelműségből származó szavak esete.

(2) REF: Különben nem hoznák haza.
RES: Különben nem hoznak haza.

A másik típus az azonos jelentésű, de különböző alakú szavak elrontása. Ezeket az eseteket szintén nem számoljuk hibásnak. Ha ezeket az eseteket nem tekintjük hibáknak, akkor a rendszerünk **99,83%-os** relatív pontosságot ér el.

(3) REF: Hova mész nyaralni?
RES: Hová mész nyaralni?

A hibáknak kevesebb mint fele valódi hiba, de a valódi hibák fele a tulajdonnevek helyesírásából fakad. Ezt azért fontos megemlíteni, mert ha a gép soha sem látott példát egy-egy tulajdonnév helyesírására, akkor nem is várhatjuk el tőle, hogy tökéletesen ékezetesítse azt.

Hibatípus	**Arány (db)**	**Példák** (referencia (ref) - eredmény (res))
Helyes kimenet	**55,07% (38 db)**	
Ekvivalens alakok	7,26%	hova - hová, tied - tiéd
Értelmes kimenet	92,74%	ref: Érdekelné ez a dolog? res: Érdekelne ez a dolog? ref: Különben nem hoznák haza. res: Különben nem hoznak haza.
Valódi hibák	**44,93% (31 db)**	
Tulajdonnevek	45,16%	Liúról - Liuról, Ramával - Rámával
Hibás értelmezés	54,84%	még - meg, melyen - mélyen, teli - téli

2. táblázat. A Tranformer modell különböző hibatípai

A Transformer modell teljesítményének egyik érdekes eredménye, hogy kizárólag csak ékezettel kapcsolatos hibákat vétett, de ez az RNN modell vagy a MOSES esetén nem így volt. A 3. táblázatban látható az RNN modell és a MOSES azon hibatípusai, amelyek eltérnek a Transformer modelltől.

Modell	Hibatípus	Példa (referencia - eredmény)
RNN, MOSES	Kis- és nagybetű	Átvehetitek - átvehetitek; Azt - azt
	Más formátum	7:54 - 7: 54; a"... - a..."; Szóval, ... - Szóval,..... Gary's - Gary 's
MOSES	Nem ékezesített	széttörnéd - szettorned; hőtermelés - hotermeles; túlárasztotta - tularasztotta

3. táblázat. Az RNN modell és MOSES hibatípusai, amelyek eltérnek a Tranformer modelltől

7. Összegzés

A kutatásunkkal létrehoztunk egy ékezetvisszaállító rendszert. A rendszer tanításához egy neruálishálózat-alapú gépi fordítórendszert használtunk, amely transzformer modellt és Sentence Piece tokenizálót használ. A rendszerünk 99,63%-os pontossággal tudja helyesen visszaállítani az ékezeteket. Végeztünk hibaelemzést is és azt állapítottuk meg, hogy a hibák közel fele nem is valódi hiba, ha ezeket az eseteket helyes kimeneteknek tekintjük, akkor a rendszerünk 99,83%-os pontosságot ér el. A rendszerünkhöz készítettünk egy demó felületet is, amelyen ki lehet próbálni a különböző modellek működését.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 számú projekt keretében az FK 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (szerk.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.0473

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., MĂźller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019), http://www.aclweb.org/anthology/W19-5301

Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation. pp. 187–197. Edinburgh, Scotland, United Kingdom (July 2011), https://kheafield.com/papers/avenue/kenlm.pdf

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 177–180. ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)

Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 66–75. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), https://www.aclweb.org/anthology/D18-2012

Mihalcea, R., Nastase, V.: Letter level learning for language independent diacritics restoration. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. pp. 1–7. COLING-02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), https://doi.org/10.3115/1118853.1118874

Nagy, P.: Magyar nyelvű zajos szövegek automatikus normalizálása. Szakdolgozat, Pázmány Péter Katolikus Egyetem (2018)

Németh, G., Zainkó, C., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P.: The design, implementation, and operation of a hungarian e-mail reader. International Journal of Speech Technology 3(3), 217–236 (Dec 2000), https://doi.org/10.1023/A:1026567216832

Novák, A., Siklósi, B.: Automatic diacritics restoration for Hungarian. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language

Processing. pp. 2286–2291. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015), https://www.aclweb.org/anthology/D15-1275

Novák, A., Siklósi, B.: Ékezetek automatikus helyreállítása magyar nyelvű szövegekben. XII. Magyar Számítógépes Nyelvészeti Konferencia pp. 49–58 (2016)

Scannell, K.P.: Statistical unicodification of african languages. Language Resources and Evaluation 45(3), 375 (Jun 2011)

Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR abs/1508.07909 (2015), http://arxiv.org/abs/1508.07909

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Yarowsky, D.: A Comparison of Corpus-Based Techniques for Restoring Accents in Spanish and French Text, pp. 99–120. Springer Netherlands, Dordrecht (1999)

Zweigenbaum, P., Grabar, N.: Accenting unknown words in a specialized language. In: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. pp. 21–28. Association for Computational Linguistics, Phildadelphia, Pennsylvania, USA (Jul 2002)

Elírások automatikus detektálása és javítása radiológiai leletek szövegében

Kicsi András[1], Szabó Ledenyi Klaudia[1], Németh Péter[1], Pusztai Péter[1,2], Vidács László[1,2], Gyimóthy Tibor[1,2]

[1]Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
Szeged, Dugonics tér 13.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
{akicsi,ledenyik,nemethp,pusztaip,lac,gyimi}@inf.u-szeged.hu

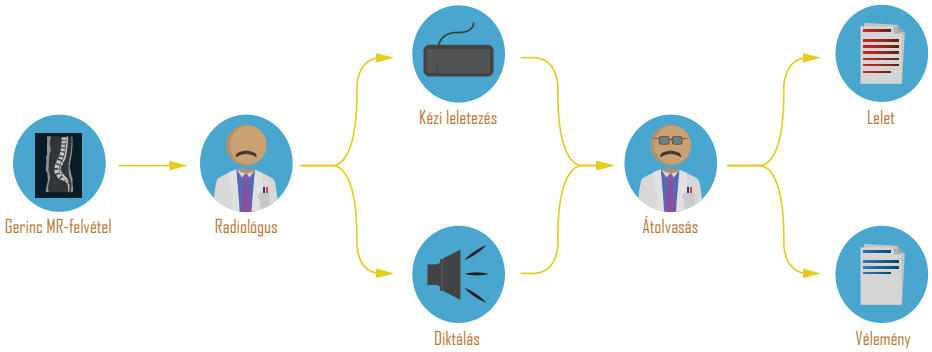
Kivonat A radiológiai leletezés közben gyakran előfordulhatnak szövegbéli hibák, melyek kézi javításra szorulnak. Ez időt von el a radiológustól, valamint a hibák sikertelen felismerése nyomán rontja a leletek minőségét és utólagos gépi feldolgozhatóságát is. Cikkünkben magyar nyelvű gerincleletek elírásainak automatikus kijavításával foglalkozunk. Ismertetjük az általunk felhasznált módszereket, és megmutatjuk, hogy az elírások automatikus javítása az értelmezést is nagymértékben javítja. Módszerünket 882 valós lelet kézi hibajavításával vetjük össze.
Kulcsszavak: radiológia, elírásjavítás, helyesírás, információkinyerés, NLP

1. Motiváció

A modern egészségügyben egyre nagyobb hangsúlyt kapnak a különböző számítógépes megoldások, ezen belül a mesterséges intelligencia is (Novák és Siklósi, 2015, 2016; Mykowiecka és Marciniak, 2007). Korszerű szoftverek képesek támogatni a műtéti tervezést, képi adatokon anomáliák keresését, és a klinikai leletezést is. A leletezésben például már egyáltalán nem ritka, a magyar egészségügyben sem, hogy az orvos gépelés helyett diktálja a leleteket egy szoftvernek, amely ezeket automatikusan rögzíti, igen jó minőségben, magyar nyelven is. Elmondható tehát, hogy rohamos fejlődés figyelhető meg az automatizálásban és az orvosok munkájának gépi segítésében (Kernighan és mtsai, 1990; Lai és mtsai, 2015).

A radiológiai vizsgálatokban is ugyanígy jelen van naprakész technológia. Jelen cikkben a diagnosztizálás ezen szegletével foglalkozunk, sőt ezen belül is csak a radiológiai gerinc vizsgálatokkal. Csak Magyarországon évente sok ezer gerinc Röntgen- és MR-felvétel készül. Ezek szöveges formában, magyar nyelven íródnak. A szakorvos elküldi a pácienst radiológiai vizsgálatra, a felvételek elkészülnek, majd ezeket egy radiológusnak továbbítják, aki szöveges lelet formájában rögzíti a képeken látott érdemleges információt. A lelet kulcsfontosságú része a radiológiai vélemény, mely összefoglalva tartalmazza a leletben tett legfontosabb megfigyeléseket, tömör orvosi szakzsargonnal, valamint latin rövidítésekkel

megfogalmazva. A lelet a radiológustól visszakerül a szakorvoshoz, aki az információ birtokában meghozza a páciens további kezelésére vonatkozó döntéseket. A leletezés folyamatát az 1. ábra illusztrálja.



1. ábra: A radiológus munkája a lelet rögzítése során

A leletezés során a radiológusnak már számos magyar klinikán is lehetősége nyílik diktáló szoftver használatára. Tradicionálisabb esetekben maga a radiológus gépel, vagy esetleg radiográfus asszisztens, szintén diktálás nyomán. Nem nehéz belátni, hogy a megfelelő minőségű diktálás értékes időt takaríthat meg. A magyar nyelvű klinikai diktálók minősége is igen fejlett már, bár valóban jó eredményeket leginkább egyéni tanítás után, és csak a speciális területen belül tudnak elérni. A radiológus számára tehát jelenleg adott a választás lehetősége, hogy időt szán a szoftver használatának betanulására, esetleg tanítására, vagy ragaszkodik a régi módszerekhez.

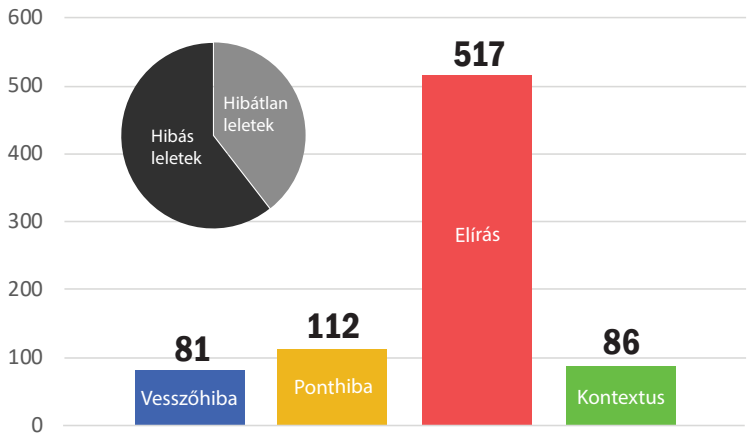
Bármelyik esetet is választja, az elkészült leletet valószínűleg át kell olvasnia a hibák kiküszöbölése érdekében. A hiba természetesen lehet kritikus, például ha kifelejtett valamit a leletből, viszont leggyakrabban csak elírásokról van szó. A diktáló szoftvereknél egyértelműen szükség van még erre a lépésre, ám pozitív oldaluk, hogy elírásokat, betűcseréket és értelmet nélkülöző szavakat sokkal kisebb valószínűséggel írnak a leletbe, hiszen értelmes szavakon tanították azokat.

A leletekben különösen gyakorinak számítanak a felcserélt betűk, amelyek latin szavakban még kevésbé észrevehetők, illetve az elmaradó ékezetek. Van például olyan radiológus, aki ugyan ír ékezeteket, az "ő" és "ű" betűket konzisztens módon ékezetmentesíti. A hibák természetesen nem csak a gépi, hanem az emberi értelmezést is ronthatják. A kézi javítás során körülbelül minden negyvenedik leletben volt olyan mondat, ami egyszerűen félbemaradt, vagy szavai szórendje alapján nem lehetett megfelelően értelmezni.

Kutatásunkban radiológiai leletek automatikus értelmezését tűztük ki célul, a témában korábban már publikáltuk módszerünket (Kicsi és mtsai, 2019), amellyel radiológiai leletekben szereplő testrészeket, elváltozásokat és tulajdonságokat detektáltuk automatizált módon, gépi tanulási módszerekkel. A testrész az

emberi test pontos részének megnevezése (például "L.V. discus"), elváltozás lehet bármely kóros elváltozás ("előboltosulás"), aspektus ("magassága") vagy normális állapotot jelző kifejezés ("normális"). A tulajdonság egy elváltozást módosító, mértéket ("3 mm-es") vagy minőséget ("körkörös") leíró szó. A tanításhoz 487 leletből álló, radiológus által annotált tanítóhalmazt alkalmaztunk. A detektálás természetesen nem jelent teljes értelmezést, publikációnk óta már rendelkezünk kialakított módszerrel a testrészek és elváltozások azonosítására, és a különböző elemek kapcsolatainak megállapítására is. Ennek pontos módszerei nem képezik jelen cikk tárgyát, ám az itt megfogalmazott célok jelentős motivációt szolgáltatnak az itt leírt kísérleteinkhez, illetve értelmezési munkánk az itt ismertetett módszer kiértékeléséhez is hozzájárulnak.

A gépi értelmezés az emberi látásmódtól lényegesen különbözik, így érthető, ha a számítógép esetleg nem képes túllépni a leletekben előforduló elírásokon, azok további hibákat indukálhatnak, amely rontja a rendszer megbízhatóságát. Természetesen a tanítóhalmaz átnézhető és kijavítható utólag, kézi átnézéssel, ahogy ezt meg is tettük. Ám a való életben történő használathoz fel kell készülnünk arra, hogy ilyen hibákat a jövőben is fognak véteni a leletet gépelő személyek. Ezért egy olyan módszert dolgoztunk ki, amellyel az elírásokat nagy valószínűséggel detektálni, és automatikusan javítani tudjuk a gerinc leletekben.



2. ábra: Kézzel talált hibák a 487 leleten

A tanítóadatként felhasználni kívánt 487 leleten első körben kézi javítást végeztünk, mely során a leletekben található összes hibát feljegyeztük a hiba helyével és a hibás mondat szövegével együtt, valamint egy alternatív helyes mondatot is megadtunk a hibás mellé. A hibákat utólag kézi átnézéssel csoportosítottuk négy halmazba. Kézenfekvő hibák a vesszőhibák, melyeket gyakorta ejthetnek mind kézi gépeléssel, mind diktálással, ez lehet kimaradó, vagy felesleges vessző egy mondatban. A ponthibák nagyon hasonlóak, itt legtöbbször a lelet végére, vagy esetleg a szövegtörzs mondatainak végére felejtett el a radio-

lógus pontot tenni. Ez utóbbi a gépi feldolgozás szempontjából lényegesebb a vesszőknél, mivel sok nyelvi elemző egy mondaton belül elemez, ezt pedig az írásjelekből állapítja meg. Nagyon gyakori hiba az elírás, amely annyit jelent, hogy önmagában értelmetlen szó került a leletbe, ez történhet betűk hozzáadásával vagy kihagyásával, felcserélt betűkkel, vagy helytelen ékezetekkel. Latin szó magyar ragozásával megengedőek voltunk, ezek ugyanis nem kiforrott szabályok alapján történnek, viszont nagyon gyakoriak a leletekben. A negyedik hibatípus a kontextusfüggő hibák, amelyek önmagukban helyes szavak helytelen használatán alapulnak, ide tartozik a rossz ragozás, illetve olyan elírások is, amelyek mégis értelmes szót produkálnak. A 487 leletből 295 tartalmazott valamilyen hibát, a hibák nagyon gyakran tömegesen jelentkeztek. A felfedett hibákról statisztikát láthatunk a 2. ábrán.

Ezen eredményekből látható, hogy a hibák nagy arányban olyan elírásokat tartalmaznak, amelyek kontextustól független módon javíthatók lennének. Természetesen egyszerű átírási szabályokkal a pontosan ilyen hibákat ezután is sikeresen tudnánk javítani, de a lehetséges elírási hibák száma óriási, így erre egy automatizált módszer kell. Cikkünkben az erre irányuló kutatásunk eredményeit mutatjuk be. Tehát jelenleg az értelmetlen szavakból és ékezethibákból adódó szavak detektálását és automatikus javítását tűztük ki célként. Az eredményeinket a leletértelmezési kutatásunkban kívánjuk felhasználni, ezért különösen fontos kérdés, hogy a hibajavítás mennyiben befolyásolhatja a névelemfelismerési és -normalizálási feladatokat. Bár a gerincleletek területe igen szűkös, mégis hisszük, hogy az jó esettanulmányként szolgálhat más detektálási vagy azonosítási feladatokhoz is. A kutatás során fejlesztett helyesírásjavító eszközt továbbá közvetlenül a leletezés folyamatában is fel lehetne használni, hogy a helyesírási hibák már keletkezésükkor javításra kerüljenek. Természetesen a hamis találatokat ehhez minimalizálni kell, azonban egy megfelelő automatikus javítás nagyban csökkentheti a leletek végső átolvasási idejét. Bevezetjük tehát az alábbi kutatási kérdéseket:

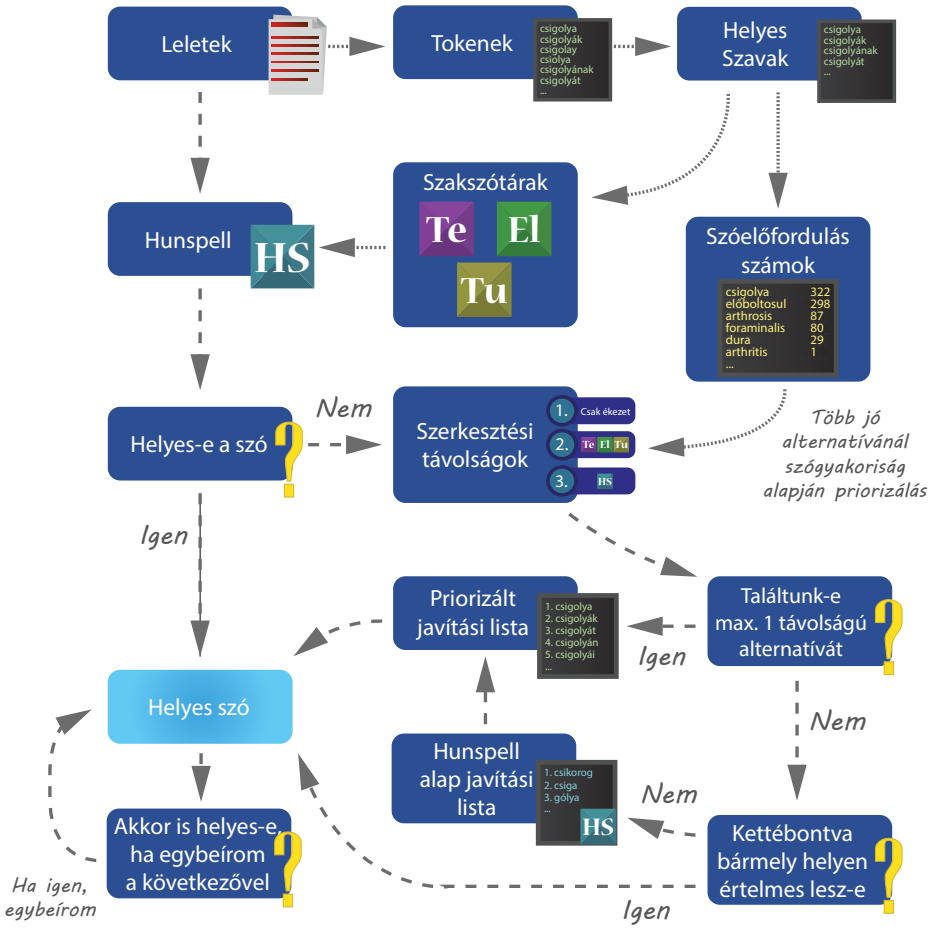
- **RQ1:** Az elírási hibák megfelelő javítása pozitívan befolyásolja-e a detektálási és azonosítási feladatok eredményét?
- **RQ2:** Mekkora részben lehet az emberi kiértékelés helyett az automatikus módszerre hagyatkozni?

2. Módszer

Jelen munkában a radiológiai gerincleletek szövegében törekszünk elírási hibák automatikus felismerésére és kijavítására. A feladatot a Hunspell[1] helyesírásellenőrző szoftver segítségével valósítottuk meg, amely nagy népszerűségnek örvend morfológiailag gazdag nyelvek szövegfeldolgozásában. A rendszer egy beépített szótárral dolgozik, és különböző ragozási és egyéb szabályokkal van ellátva, amelyeknek köszönhetően remekül szűr általános témájú szövegben. A Hunspellt önmagában kipróbálva, irreálisan nagyszámú hibát jelölt a leletekben, ugyanis

[1] http://hunspell.github.io

rengeteg latin, és egyéb orvosi szót hibásnak jelölt, mivel ezek nem voltak benne a szótárában. Természetesen ez nem róható fel a rendszernek, hiszen magyar nyelvű szövegre van kialakítva, a leleteknél pedig az egyszerű latin szavak mellett nagyon gyakran előfordulnak olyan, helyesnek tekintendő szavak is, ahol egy latin szó magyar ragozással szerepel, mint például „herniatiora” vagy „protrusiojának”. Ezen szavak a leletek értelmezhetőségét nem rontják, hiszen az orvosok ezeket konzisztensen és rendszeresen használják.



3. ábra: Automatikus elírásjavító módszerünk működése

Kifejlesztett módszerünk tehát a Hunspell rendszerre támaszkodik, ám szakszótárakat és priorizálási szabályokat kellett kiépítenünk a gerincleletek megfelelő kezeléséhez. Módszerünk teljes sémáját a 3. ábrán mutatjuk be. 5649 lelet állt

összesen rendelkezésünkre, ebből 487 radiológus által beannotált, tanulóadatnak is alkalmas lelet. A helyesírás fejlesztéséhez azonban további leleteket is kiértékeltünk, így összesen 882 leletet használtunk kiterjedt szótárak létrehozásának reményében. A Hunspell eredeti szótára sajnos nehezen birkózik meg az orvosi, nem ritkán latinul írt szakkifejezésekkel. Az ábrán elsőként a pontozott vonalak mentén indulunk el. Az 5649 leletekből első lépésben kinyertük az összes testrészként, elváltozásként vagy tulajdonságként annotált tokent (a Bi-LSTM (Hochreiter és Schmidhuber, 1997) technikával történő detektálás a szövegkörnyezet alapján, olyan szavakat is képes helyesen felcímkézni, amelyeket még nem látott). A Bi-LSTM rendszer által felismert névelemeket lexikografikus rendezés után egyesével átnéztük, a hibás szavakat pedig kézzel szűrtük. Igyekeztünk körültekintően eljárni. Mivel nem rendelkezünk a szükséges orvosi háttértudással minden szó pontos értelmezéséhez, így külön gyűjtöttük az általunk dilemmásnak ítélt tokeneket, amelyekre gyanakodtunk, hogy tartalmazhatnak elírást, ám orvosi szaknyelvhez tartoznak. Ezen dilemmás eseteket a radiológus kézi ellenőrzésével oldottuk fel. A helyes szavakat ezután három szakszótárba soroltuk (testrész, elváltozás és tulajdonság), melyek összesen 5723 szót tartalmaznak jelen állapotban. Ugyanezen helyes szavakat nem csak a szakszótárakban használjuk fel, hanem később, a javítások priorizálásánál is. Ezért az összes szóhoz megállapítunk egy gyakoriság listát, amely aszerint van rendezve, hogy hányszor fordultak elő az összes leletben, ennek felhasználását később látjuk. A szakszótárakat a Hunspell rendelkezésére bocsájtjuk, amely az adatok birtokában bírálja el egy adott új szó helyességét.

Amennyiben helyes volt a szó, akkor nincs igazán teendőnk. Ilyenkor kísérletet teszünk két szó összevonására, mivel gyakori például a „csigolyatest” szó leletbéli különírása „csigolya” és „test” szavakként, amik ugyan külön is értelmesek, de a radiológus valószínűleg egyben szándékozott leírni. Tehát minden helyes szóra leellenőrizzük azt is, hogy ha a következő szóval egybeírnánk, akkor is értelmes szót kapnánk-e. Ha igen, akkor összeolvasztjuk a két szót.

Ha a szó helytelen volt, azaz nem található meg sem a Hunspell saját szótárába, sem pedig a szakszótárakban, akkor megpróbáljuk kijavítani azt. A Hunspell alapértelmezetten egy szerkesztési távolságon alapuló priorizált listát ad a lehetséges javításokról. Ez azonban nem volt megfelelő, szintén a speciális szavak miatt. A priorizálást átalakítottuk olyan módon, hogy először az ékezethibákat tartsa valószínűnek, például az „eloboltosulás” szónál az „előboltosulás” előrébb kerül a listában mint az „elboltosulás”. Amennyiben ilyen alternatívát nem talált, akkor az egyszerű szerkesztési távolságot veszi alapul. Először a szakszótárak szavai közt, majd a saját beépített szótárának szavai között keres javítási lehetőséget. Ezen belül tovább rangsorol a szógyakoriságok alapján, amelynek listáját korábban állítottuk elő. Ez alkalmas az olyan esetekre, amikor két értelmes alternatíva is van, amelyet néha még kézi vizsgálattal is csak nagy odafigyeléssel, vagy tudással lehet javítani. Ilyen például az „arthrotis” szó, amely egyaránt kijavítható „arthritis” és „arthrosis” irányba. Ilyenkor a gyakoribbat választjuk, ami a gerinc területén egyértelműen az „arthrosis” lenne.

Ezután ellenőrizzük, hogy találtunk-e olyan javítást, ami maximum (vagyis valójában pontosan) 1 szerkesztési távolságra van az eredeti szótól. Ezek a legvalószínűbb elírások, ugyanis itt csak egy betű hiánya vagy félreütése jelentkezett. Ha találtunk ilyet, akkor azok egyenesen mennek tovább a priorizálásukat megtartva a javítási listába, majd közülük a legvalószínűbb lesz az ajánlott automatikus javítás. Ha nem volt ilyen alternatíva, akkor ahelyett, hogy valami nagyon különböző szót ajánlanánk, először megpróbáljuk a szót felbontani két, vagy ha ez nem sikeres, akkor akár három részre, ahogy például az „előboltosulódiscus" hibás szó felbontható „előboltosuló" és „discus" helyes szavakra. Ha ezután a szavak mindegyike értelmes, akkor automatikusan azt tekintjük helyes javításnak. Ez természetesen hordoz veszélyeket például szóeleji vagy szóvégi „a" betűknél, de a tapasztalatok alapján nem okoz komoly hibákat. Ha a felbontás sem sikeres, akkor a szerkesztési távolságokkal előállított priorizált listát adjuk tovább, ahogy azt pár lépéssel korábban előállítottuk Hunspell segítségével.

Az esetek többségében tehát egy valamilyen módon priorizált listát kapunk. Az elkészült módszer tapasztalataink szerint szinte soha nem ad hamis riasztást, illetve az ajánlások igen pontosak, az algoritmust leginkább csak a ragozás tévesztheti meg. A megvalósított rendszer futásideje igazán rövid, tulajdonképpen észrevehetetlen, bőven egy másodpercen belül teljesít egy egész leletre, így alkalmas lehet akár írás közbeni ellenőrzésre és javításra is.

3. Eredmények

Motivációnk bevezetésekor két kérdést tettünk fel. Ezek azt vizsgálják mennyiben javítja elírásjavító módszerünk a leletek automatikus értelmezésének minőségét, illetve hogy mennyivel rosszabb, vagy esetleg jobb elírásjavítást kapunk így, mintha emberi munkával, kézileg néznénk át a leleteket.

A leletek értelmezésével foglalkozó kutatásunk (Kicsi és mtsai, 2019) kapcsán képesek vagyunk testrészek, elváltozások és tulajdonságok detektálására, és pontos azonosítására is. Ezen eredmények nem képezik jelen cikk tárgyát, ám illusztrálhatják, hogy a gépi értelmezésre milyen mértékben és irányban hathat ki helyesírásjavító módszerünk. A detektálás Bi-LSTM technikával történt, a modellt 487 annotált leleten tanítottuk, melyből 70% képezte a tanító, 10% a validációs, a maradék 20% pedig a teszthalmazt. A rendszer nyelvi jellemzőket is felhasznál, amelyeket a magyarlánc (Zsibrita és mtsai, 2013) nyelvi elemző eszközzel nyertünk ki a szövegből. Azonosítási módszerünk szintén nyelvi elemzést, valamint szabály alapú módszereket használ testrészek és elváltozások azonosítására egy saját azonosítóhalmaz alapján. A tulajdonságokkal, azok óriási változatossága miatt nem foglalkoztunk.

Felmerül a kérdés, hogy a detektálási eredményekre kihathatnak-e az elírások. Természetesen kihathatnak, hiszen az elírások azon felül, hogy más szóalakot eredményeznek, a magyarlánc elemzését is megzavarják. Több esetben megfigyeltük, hogy az elírt szavakhoz a magyarlánc nem a megfelelő nyelvi jellemzőket rendeli és mivel ezek a Bi-LSTM modell tanítóadatának szerves részét képezik jelentős mértékben torzíthatják a tanítóadatot. A lemmatizálás lényege,

hogy egységes alakra hozza a ragozott szavakat, mely jelentősen tudja javítani a tanítási eredményeket. Ugyanezen gondolatmenetet alkalmazva, az ékezethibákból fakadó számos különböző szóalak egységesítése (javítás által) szintén segítheti a tanítás eredményességét. Detektálási eredményeinket egy mikro-átlaggal összegezve kezdetben 91,09 F1-mérték eredményt kaptunk. Az elírások javítására kidolgozott automatikus módszerünk futtatása és újabb tanítás után az eredményünk már 91,52 F1-mérték, amely csaknem fél százalékos javulás.

Az azonosítás eredményeire még valószínűbb, hogy kihat a javítás, hiszen itt szavakat vagy szótöveket próbálunk szabály alapon illeszteni a szövegre, ha pedig a szövegben ezek nem megszokott formában szerepelnek, akkor valószínűleg egyáltalán nem tudjuk automatikusan azonosítani őket. Itt szembetűnő minőségjelző szám lehet azoknak a testrészeknek és elváltozásoknak a száma, amelyeket egyáltalán nem tudtunk azonosítani. Az eredeti adathalmazban ezek az azonosíthatatlan elemek 6358 testrészből 505 testrész, és 7794 elváltozás közül 521 elváltozás volt. Az elírások javítását követően ezek az értékek 488 testrészre és 332 elváltozásra módosultak. Látható tehát, hogy javításunk valóban nagymértékben kihatott az azonosítási feladat sikerére.

RQ1 válasz: Vizsgálataink alapján a detektálást közepesen jelentős mértékben (saját rendszerünknél 0,43%-ban) pozitívan, míg az azonosítást jelentős mértékben (saját rendszerünknél több mint 20%-ban) pozitívan befolyásolja az általunk felvázolt automatikus elírásjavító módszer.

Második kérdésünk megválaszolásához a kézi adatokra támaszkodtunk. Ezeket, ahogy azt a motivációnál leírtuk, a 487 lelet kézi átnézésével, hibák és javításaik rögzítésével és utólagos klasszifikációjával állítottuk elő. A pontosabb kiértékeléshez ezt bővítjük további 395 lelet ugyanilyen módszerű kézi átnézési eredményeivel, így kiértékelésünket összesen 882 lelet szövegén végezzük. A további leleteket véletlenszerűen válogattuk. Természetesen továbbra is csak az elírásokat tűzzük ki célul, így a vesszőhibákkal nem foglalkozunk. A kézi hibajavítás során összesen 908 esetben történt elírásnak címkézett javítás. A gépi módszernél ez a szám 1248 volt. A kézi hibakeresést minden esetben ugyanaz a személy végezte, a találatok pedig másik személy által kerültek ellenőrzésre és klasszifikációra. Már első ránézésre látható, hogy a gépi hibakeresés jóval több javítást produkál. Kézzel kiértékeltük a két módszer különbségét.

A kézi adathalmazon 36 olyan javítás volt, amelyet nem jelölt a rendszer, ezek többsége értelmes szót adott a hibával, vagy csak hiányzó pont vagy kötőjel volt benne („iv”,„kb 3 mm”). Több esetben volt felesleges javítás is, mint például a „folyadéktartalmú” szó szeparálása. Az ellenőrzés során 5 olyan esetet találtunk, amely valóban hibás volt, és rendszerünk nem fedte fel, ezek a szótár szűrésének tökéletlenségéből eredhetnek. 17 esetben ugyanarra a szóra különböző javítást adott a kézi keresés és rendszerünk, ebből több valójában ekvivalens („levő”,”lévő” vagy „normál”,”normális”), és itt is volt több rossz emberi javítás vagy kontextusfüggő javítás, amelyet nem lehet kitalálni a mondatkörnyezet nélkül, például ragozási forma. 7 olyan eltérés volt, amelyet valóban hibás javításnak ítéltünk meg. Ezek leginkább olyan esetek, amikor a névelő egyben volt a szóval, a javításunkban pedig teljesen kimaradt („Abal”).

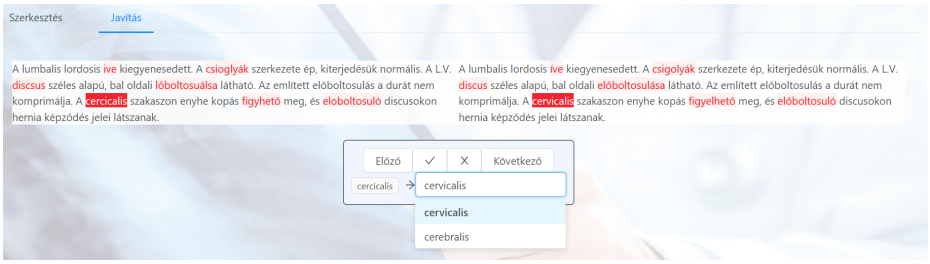
További nagyon lényeges kérdés, hogy mennyire valós a 328 hiba, amit az automatikus keresés felfedett, az emberi javító pedig nem. A hibajavítás monoton feladat, és betűcserék felett könnyen elsiklik az emberi szemlélő, így elképzelhető ekkora kihagyás. Ennek vizsgálata során a következőket állapítottuk meg: a detektált elemekből összesen 4 javítás volt indokolatlan, ezek önmagukban nem szokványos, de valószínűleg szándékosan így írt szavak („VA”,”VB”, ”radici”,”gerinccsatorna- és”). További 3 szó indokoltan volt jelölve, ezeket azonban rossz javítással láttuk el. Tehát megállapíthatjuk, hogy a 328 detektált hibából 324 emberi szemlélő számára is indokolt hibajavításnak tekinthető. Belátjuk tehát, hogy rendszerünk valóban képes emberi javítással összemérhető, sőt, túl is szárnyaló hibajavításra. A számok közti jelentős különbség alapján felmerülhet, hogy ha 1244 közeli javítás valóban indokolt, mennyire tekinthetjük a kézi kiértékelés eredményét egy megbízható összehasonlítási alapnak a nagyszámú kihagyott javításával. A kézi kiértékelés természetesen tökéletlen, de ez pusztán az emberi figyelem, és nem a kiértékelő személy hiányosságait tükrözi. Az is ezt mutatja, hogy egy orvos a jelen leleteket egyszer már át kellett, hogy olvassa, az említett számos hiba azonban így is előfordul.

Mivel a kézi javítás eredményeit felhasználva készítettük módszerünket, nem meglepő, hogy ezeken jól teljesít, hiszen rájuk optimalizáltunk. Ezért további 300, nem optimalizált lelettel is kipróbáltuk ugyanezt, az eredményeket kézzel vizsgáltuk át ismét. A kísérletben 441 hibariasztást kaptunk. Ezekből az átvizsgálás során azt szűrtük le, hogy 11 esetben történt indokolatlan jelölés, és további 6 esetet indokoltan, de rosszul javított. A szótárak bővítésével természetesen ezen hibás találatok és hibák nagy része is kiküszöbölhető lenne.

RQ2 válasz: Szűk területünkön belül az automatikus javítás jelentősen, több mint 30%-al több valós hibát tárt fel a kézi ellenőrzésnél, a javítás minőségének különösebb romlása nélkül.

Végezetül módszerünk használhatóságát egy felületen is demonstráljuk. Ennek képernyőképe látható a 4. ábrán. A felületen bal oldalon egy kitalált lelet számos elírást tartalmazó szövege látható. Ezeket, amint az ábrán is látható, módszerünk helyesen detektálta, piros színnel jelölte ki. A jobb oldali ablakban láthatjuk a módszerünk által összeállított automatikus javítást, az ábrán ezek mindegyike helyes javítás. Amennyiben hibás javítást tapasztalnánk, a javításokat az alsó panelen tudjuk elbírálni. Egy javítást elfogadhatunk, elvethetünk amennyiben a hiba nem volt valós, vagy akár módosíthatunk is. Ez utóbbihoz nagyban hozzájárul, hogy módszerünk egy priorizált listát szolgáltat, amelyre szintén láthatunk példát az ábrán. Amennyiben a listában sem találjuk meg a kívánt javítást, abban az esetben újat is beírhatunk a szövegdobozba. Az ábrán szereplő javaslatok módosítás nélkül, automatikus javításként álltak elő.

A felület jelenleg a módszer hibáinak javítása szempontjából hasznos, hiszen könnyen kipróbálhatók benne különleges megfogalmazású szóalakok is. A jövőben azonban akár hibák gyors kézi annotációjában is segíthet, elég adatot kinyerve megtalálhatjuk a tipikusan nem kezelt hibákat, bővíthetjük a szótárakat, illetve akár még egy esetleges gépi tanulási módszer bemenetét is alkothatják.



4. ábra: Automatikus helyesírásjavító felületünk radiológiai gerinc leletek ellenőrzésére: Bal oldal: eredeti szöveg, Jobb oldal: javítások, Alsó panel: kezelés további lehetséges találatokkal

Fontos megjegyezni, hogy módszerünk kizárólag magyar nyelvű radiológiai gerinc leletek elírásjavítására lett optimalizálva. Az nem jelenthető ki, hogy más területeken is garantáltak lennének a hasonló eredmények, vagy a módszer sikere. Jelen területünk viszonylag szűkös szókészlettel dolgozó orvosi szakszöveg. Feltételezzük, hogy hasonló, behatárolható orvosi területeken új szótárakkal hasonló eredmények produkálhatók, ám az általános szöveg elemzésében már sokkal rosszabbul teljesít megközelítésünk, hiszen szakszavakat priorizál. Az általános elírásjavítókkal szemben azonban nagy előny, hogy valóban támaszkodhatunk a terület szokásaira, ezzel sokkal pontosabb eredményeket adva a területen. Motiváló példaként kipróbáltuk, hogy egy ismert szövegszerkesztő magyar nyelvű ellenőrzője a 4. ábra szövegében ugyan detektál minden hibásan írt szót, de ezek egyikére sem adja meg a helyes javítást, valamint 5 további szót is hibásnak jelöl, amelyek a leletekben azonban nagyon gyakran előfordulnak.

4. Kapcsolódó kutatások

Az egészségügyi rendszerben a leletezés túlnyomórészt szabad megfogalmazású formában történik. Ez egyfelől hasznos a beteg állapotának szabatos és pontos leírása szempontjából, másfelől viszont jelentősen megnehezíti a leletekből történő információkinyerést. A szabad megfogalmazású leletek használatának egy további hátránya a bonyolult információkinyerésen kívül, hogy ez a leletezési forma sokkal több hibalehetőséget rejt magában a strukturált leletezéssel szemben. Hibák alatt itt elsősorban elgépelésből, vagy helytelen hangfelismerésből származó elírási hibákra kell gondolni. Egy-két elírt szó a leletben nem tűnik nagy problémának, azonban ha az az elírás pont egy kulcsfontosságú részen történik, a diagnózist vagy esetleg a kezelési formát érintve, akkor az komoly következményekkel járhat a páciens egészsége szempontjából. Mindezek mellett információkinyerés szempontjából is jelentőséggel bírnak az elírt szavak, ugyanis több rendszernél is bevett eljárás a szavak ontológiákhoz történő hozzárendelése és kódolása, ez pedig pontos szóegyezéses alapon működik elsősorban, így egy elírt szó információvesztést jelent (Tolentino és mtsai, 2007).

Kukich hibajavításokkal kapcsolatos átfogó tanulmányában (Kukich, 1992) három fő elírási kategóriát állapított meg, melyekkel a hibajavító rendszereknek meg kell birkózniuk: nem létező szó felismerése (nonword error detection), izolált szójavítás (isolated word error correction), kontextusfűggő hibajavítás. A nem létező szó felismerési technikák két fő csoportja az n-gram analízisen alapuló módszer, mely során szokatlan karaktermintázatok alapján becsülik a hiba lehetőségét, illetve a szótár alapú módszer, mely során egy nagy szótárral való egyezéskeresés történik. Az izolált szójavító algoritmusok többsége valamilyen szerkesztési távolság alapján ajánl javításokat. Egy tanulmány szerint az elírások 80%-át a következő hibák teszik ki: szóhoz hozzáírt betű, szóból kihagyott betű, a szó egy betűjének más betűvel történő helyettesítése, illetve a szó két betűjének felcserélése (Damerau, 1964). A kontextusfüggő hibák esetén egy helyesen írt szó egy másik helyesen írt, nem oda illő szóval van helyettesítve. Az ilyen hibák szűrésére statisztikai alapú nyelvi modelleket alkalmaznak elsősorban.

A nemzetközi szakirodalomban több ízben is találhatunk példákat orvosi szöveg javítására, így például Tolentino és szerzőtársai vakcinabiztonságról szóló jelentésekben végeztek hibajavítást, mely során szótárat készítettek általános angol szavak és területspecifikus szavak gyűjteményéből, majd egy szerkesztési távolságon alapuló módszert alkalmaztak a hibák szűrésére (Tolentino és mtsai, 2007). Crowell és munkatársai egy szabad forrású szoftvert alkalmaztak egy orvosi portálhoz intézett lekérdezésekben található elírások szűrésére, szerintük a szavak között statisztikailag kimutatható, hogy melyiket milyen gyakorisággal használják keresésre a felhasználók. Ennek alapján újrarendezték a hibajavító szoftver javaslatait a szavak keresőmotorba írásának gyakorisága alapján, ennek hatására jelentős javulást tapasztaltak módszerük pontosságában (Crowell és mtsai, 2004). Mykowiecka és Marciniak bigram alapú nyelvi modellt használtak lengyel nyelvű mammográfiai leletek automatikus helyesírás-javítására (Mykowiecka és Marciniak, 2007). A szerzők saját maguk építettek szótárt, az elírt szavak több, mint 90%-át pontosan tudták javítani, azonban a szótárukban nem szereplő, helyesen írt szavak több, mint felét az algoritmus helytelenül módosította. Patrick és szerzőtársai egy trigram nyelvi modellel rangsorolt, szerkesztési távolságon alapuló rendszert dolgoztak ki kórházi leletekben található elírások javítására. A modellhez a szótárat külső forrásokból, illetve javított leletek szövegéből építették (Patrick és mtsai, 2010). Ruch és szerzőtársai névelemfelismerésen alapuló rendszert fejlesztettek francia nyelvű leletekben található elírások javítására (Ruch és mtsai, 2003). Megvalósításukban a névelemfelismerő rendszer megtalálta a szövegben előforduló neveket, melyeket a hibajavító algoritmus figyelmen kívül hagyott. Módszerükkel a szerzők a hibás pozitív detektálásokat jelentős mértékben csökkentették (~23%-ról ~3%-ra). Kenneth és szerzőtársai noisy channel módszeren (Kernighan és mtsai, 1990) alapuló hibajavító algoritmust alkalmaztak orvosi leletekben, receptekben és allergiára vonatkozó bejegyzésekben található elírások javítására. A programhoz kiterjedt szótárat építettek változatos forrásokból, valamint névelemfelismerő rendszer segítségével a helyesen írt nevek hibás javítását is kiküszöbölték (Lai és mtsai, 2015).

A magyar nyelvű klinikai szövegekben tapasztalható elírások kezelésére korábban már folytak kutatások. Siklósi és szerzőtársai több ízben fejlesztettek klinikai szövegek rendezésére, illetve a szövegekben tapasztalható elírások automatikus javítására rendszert, mely a javítás során figyelembe vette a szövegkörnyezetet, valamint képes volt az egybeírások kezelésére is (Siklósi és mtsai, 2013a,b, 2012). A szerzők egy másik kutatásukban ékezet nélküli szavak ékezetesítésére fejlesztettek statisztikai gépi fordításon alapuló rendszert, mely az esetek 99%-ában helyes ékezetes alternatívát javasolt (Novák és Siklósi, 2015, 2016).

Az általunk fejlesztett hibajavító algoritmus a javaslatok kialakításához több alegységre támaszkodik. Az egyik ilyen egység a szabadforrású Hunspell helyesírásjavító szoftver, melyet számos népszerű nemzetközi IT cég használ rendszereiben (Firefox, Chrome, Libre Office, Photoshop, Eclipse), valamint akadémiai berkeken belül is előszeretettel alkalmazzák, változatos nyelvekre adaptálva, mint például angol (Bettenburg és mtsai, 2011; Al-Hussaini, 2017), arab (Zerrouki és Balla, 2009), vagy esperanto (Blahuš, 2009).

Kifejezetten radiológiai gerincleletekre optimalizált elírásjavító tudomásunk szerint nem született még, sem magyar, sem külföldi forrásoktól. Módszerünk a terület szűkös szókészletét és specifikusságát kihasználva javít elírásokat. Munkánk esettanulmányként szolgálhat más hasonló területek javításai számára is, nem kizárólag klinikai környezetben.

5. Összegzés

Írásunkban elírásjavító módszerünket mutattuk be, amely a Hunspell elemzővel és radiológus részvételével valós leletekből összeállított szakszótárak segítségével azonosít elírásokat magyar nyelvű radiológiai gerincleletek szövegében, és automatikus javítást képes megvalósítani. Bemutattuk technikánk lépéseit, majd beláttuk, hogy az automatikus elírásjavítás jelentős mértékben pozitívan befolyásloja a szöveg gépi értelmezhetőségét, saját detektálási módszerünkön 0.43% javulást, azonosításunkban pedig a korábban azonosíthatatlan entitások több mint 20%-át sikerült azonosítani. Eredményeinket 882 leleten értékeltük ki, kiértékelésünk alapján módszerünk 38%-al több hibát tártunk fel, javításunk pedig csak néhány esetben volt téves. A módszer kézenfekvő jövőbeli felhasználása a leletek gépi értelmezésének javítása, de a rendszer akár a leletezés során, valós idejű javítások megvalósítására is alkalmas lehet, ezzel gyorsítva a leletek átnézési idejét és javítva azok végső minőségét. Bemutattunk egy felületet, amely alkalmas jellemző hiányosságok feltérképezésére is, csakúgy mint a hibák gyors annotációjára. Ezzel a jövőben a technika tovább finomítható.

Köszönetnyilvánítás

Jelen kutatás az Innovációs és Technológiai Minisztérium ÚNKP-19-3 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült. A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM). Készült az EFOP-3.6.3-VEKOP-16-2017-00002 támogatásával.

Hivatkozások

Al-Hussaini, L.: Experience: Insights into the benchmarking data of hunspell and aspell spell checkers. J. Data and Information Quality 8(3-4), 13:1–13:10 (jun 2017)

Bettenburg, N., Adams, B., Hassan, A.E., Smidt, M.: A lightweight approach to uncover technical artifacts in unstructured data. In: 2011 IEEE 19th International Conference on Program Comprehension. pp. 185–188 (June 2011)

Blahuš, M.: Morphology-aware spell-checking dictionary for esperanto. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. pp. 3–8. Masaryk University, Brno (2009)

Crowell, J., Zeng-Treitler, Q., Ngo, L., Lacroix, E.M.: A frequency-based technique to improve the spelling suggestion rank in medical queries. Journal of the American Medical Informatics Association : JAMIA 11, 179–85 (05 2004)

Damerau, F.: A technique for computer detection and correction of spelling errors. Commun. ACM 7, 171–176 (03 1964)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997)

Kernighan, M., Church, K., Gale, W.: A spelling correction program based on a noisy channel model. pp. 205–210 (01 1990)

Kicsi, A., Pusztai, P., Szabó Ledenyi, K., Szabó, E., Berend, G., Vincze, V., Vidács, L.: Információkinyerés magyar nyelvű gerinc mr leletekből. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). p. 177–186. Szeged (2019)

Kukich, K.: Techniques for automatically correcting words in text. ACM Comput. Surv. 24, 377–439 (12 1992)

Lai, K., Topaz, M., Goss, F., Zhou, L.: Automated misspelling detection and correction in clinical free-text records. Journal of Biomedical Informatics 55 (04 2015)

Mykowiecka, A., Marciniak, M.: Domain–Driven Automatic Spelling Correction for Mammography Reports, vol. 35, pp. 521–530 (04 2007)

Novák, A., Siklósi, B.: Automatic diacritics restoration for hungarian. In: EMNLP. p. 2286–2291. The Association for Computational Linguistics, The Association for Computational Linguistics (2015)

Novák, A., Siklósi, B.: Ékezetek automatikus helyreállítása magyar nyelvű szövegekben, p. 49–58. Szegedi Tudományegyetem, Szeged (2016)

Patrick, J., Sabbagh, M., Jain, S., Zheng, H.: Spelling correction in clinical notes with emphasis on first suggestion accuracy. In: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining. pp. 1–8 (2010)

Ruch, P., Baud, R., Geissbühler, A.: Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial intelligence in medicine 29, 169–84 (09 2003)

Siklósi, B., Novák, A., Prószéky, G.: Context-Aware Correction of Spelling Errors in Hungarian Medical Documents, p. 248–259. No. Lecture Notes in Computer Science 7978, Springer Berlin Heidelberg (2013a)

Siklósi, B., Novák, A., Prószéky, G.: Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. p. 148–158. Szegedi Tudományegyetem, Szeged (2013b)

Siklósi, B., Orosz, G., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for hungarian clinical records. In: 8th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-resourced Languages. p. 29–34 (2012)

Tolentino, H., Matters, M., Walop, W., Law, B., Tong, W., Liu, F., Fontelo, P., Kohl, K., Payne, D.: A umls-based spell checker for natural language processing in vaccine safety. BMC medical informatics and decision making 7, 3 (02 2007)

Zerrouki, T., Balla, A.: Implementation of infixes and circumfixes in the spellcheckers. In: In Proceedings of the Second International Conference on Arabic Language Resources and Tools. pp. 61–65 (2009)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. http://rgai.inf.u-szeged.hu/index.php?lang=en&page=magyarlanc (2013)

Szösszenet az elveszett morfémákért
Az alaki analógiák haszna

Naszódi Mátyás

MorphoLogic KFT.
naszodim@morphologic.hu

Kivonat A jelenlegi morfológiai elemzők gyakorlati okok miatt elég pragmatikus módon készültek. A céljuk, aránylag kis munkával fedjék le a magyar nyelvű szövegeinek szóalakjait minél kevesebb hibával. Ha a célt elérték, a szabályszerű eseteket jól leírták, a deviáns, kisebb gyakorisággal előforduló eseteket kivételként, egyedileg kezelik. A vizsgálataim szerint sokkal kevesebb kivétel van. A szavak végződése szerinti csoportosítással felderíthetők azok a szavak közötti összefüggések, melyek a korábbi adatbázisokból hiányoznak. A módszer segítségével elfeledett vagy csak leíró nyelvészek által említett szógyökök, toldalékok kerülnek napvilágra. Sőt a feltárás eredményeként pontosíthatóak a praktikus célra készült nyelvészeti, nyelvi tárak.
Kulcsszavak: morfológia, lexikográfia, helyesírás-ellenőrzés

1. Bevezető

Munkám során, melyek többsége nyelvi eszközök készítése, esetenként észleltem, hogy a meglevő szavakra vonatkozó átfogó adatbázisok hiányosak abban az értelemben, hogy az adatokból kimaradnak olyan – a cél érdekében hasznosítható – információk, melyeket könnyedén lehetne részévé tenni az adatbázisnak. Jelenleg csak morfológiai rendszerekre szorítkozom, ami persze érinti a szóelemzőket, generátorokat, helyesírás-ellenőrzőket.

A 90-es évek elején Elekfi László ragozási szótárából kiindulva készíttettem helyesírás-ellenőrzőt. Már akkor is igyekeztem egységesebb, tömörebb formára hozni a nyelvészeti adatokat. Eseti megoldásim voltak, de látom, alaposabb kutatást érdemel a téma. Most, ha nem is a teljesség igényével, de módszerem megmutatásával prezentálom, milyen kis munkával milyen eredményeket lehet elérni. A módszer pedig egyszerű: tisztán formai kérdéseket teszek fel nyelvi adatbázisoknak, majd a válasz után generatív módon tesztelem, vajon igaz-e a feltételezésem. Ily módon számtalan morfémát, nem használt toldalékolási formát, szabályt fedek fel, melyek – ha visszakerülnek a lekérdezett nyelvi adatbázisba – gazdagítják, pontosítják azt.

2. Szótani adatbázisok

A szótani adatbázisok több célt szolgálhatnak. A többnyelvű szótárakkal most nem foglalkozom, mert az adott problémától távol állnak. Engem azok a szótárak

izgatnak, melyekben toldalékolásra van több-kevesebb információ. Ezek alapja – akár tagadják a készítői, akár nem – a régi Magyar értelmező kéziszótár(MTA, 1972). Ebből indult ki Papp Ferenc a tergo szótára(Papp, 1969), az összes helyesírási adatbázis(Proszéki és Kornai, 2017), de a helyesírási lexikonoknak is ez az alapja.

A legtöbb formális toldalékolási információval a helyesírás-ellenőrzők számára készített forráskódok rendelkeznek, illetve Elekfi László ragozási szótára(Elekfi, 1994), mely ugyan nem számítógépes felhasználásra készült, de precízsége miatt alkalmas arra, illetve használtam is ebből a célból. Ezek a szótárak, mint adatbázisok, három részre bonthatók.

- Szószedet: a szótövek tárháza
- Toldaléktár: a toldalékok, előtagok jegyzéke
- Toldalékolási szabályok: itt derül ki formális leírással, hogyan és milyen formában kapcsolódhatnak a morfémák.

Ez utóbbi többnyire az előző két tárban levő morfémákra akasztott jegyek alapján működik.[1]

Az adatbázisok többsége nyilvános, könyv vagy elektronikus formában elérhetők, illetve nekem ennél többhöz van hozzáférésem. A tárak elsősorban a leírás formájában térnek el. A formai különbségen túl az igei és névszói ragozásban, jelek használatában nincs lényeges eltérés. Apróbb részletekben itt is vannak különbségek: Elekfi szótára túl szigorú. Például egyes szavaknál letiltással olyan ragokat nem enged meg, melyek a gyakorlatban mégis előfordulnak. Az esetragok halmaza is különböző a különböző helyesírás-ellenőrzőkben. A ritkább régies ragozási formák nem mindegyikben szerepelnek. Ezek apróságok. Több-kevesebb munkával egy fedél alá lehet hozni az eltérő formalizmusokat. Ezzel most nem foglalkozom, bár itt is vannak olyan kérdések, amelyek a módszeremmel tisztázhatók.

A képzők és a szóösszetételek mutatnak nagy különbséget az adatbázisokban. És itt van a legnagyobb pontatlanság, mert sokszor nincs is pontos nyelvészeti leírás, csak közelítésen, megérzésen alapuló szabályok.

3. A szótövek

A régi magyar értelmező kéziszótár tartalma mintegy 70 000 tétel. Ez azt jelenti, hogy ennyi szónak magyarázza a jelentését. Ezek között vannak összetett, képzett szavak, igekötős igék és kis mennyiségben kifejezések. Ha azt nézzük, hogy ezt a szókészletet hány szótővel lehet lefedni, meglepő eredményt kapunk. Nincs 20 000. Azért nem mondok pontos számot, mert függ attól, milyen szóképzéseket, -összetételeket kezel a morfológiai rendszer algoritmikusan. Persze a jelentés

[1] Van ettől eltérő rendszer. Például a Kimmo–Koskenniemmi által jegyzett TLFA kétszintű morfológia(Koskenniemi, 1983), melynek egyik magyar adatbázisa a XEROX tulajdona, de a kiindulási alapot a MorphoLogic szolgáltatta. Ezzel nem foglalkozom, mert nem hozzáférhető az adatbázis.

nem mindig értelmezhető a morfémák jelentésének összekapcsolásával, emiatt indokolt lehet a 70 000 tétel.

Ha viszont a jelentés (fordítás) nem érdekel, akkor felesleges ekkora tár, hisz a többi alak generatív módon megkapható egy kisebb halmazból.

4. Egyéb morfémák

A ragokkal, jelekkel, mint írtam, nincs igazán gond. Jeles nyelvészek, Papp Ferenc, Seregy Lajos, Elekfi László, valamint a helyesírás-ellenőrzők alkotói ezt elég jól feltérképezték, legfeljebb elkódoltak egy-egy tételt. A képzők viszont várnak még hasonló alapmunkára. Tisztázatlan például az igék műveltetésének *-tat, -tet, -at, -et* fonetikai besorolása. Az világos, hogy mikor magas, mikor mély hangrendű egy szó, de hogy mikor kell a bevezető *-t*, és mikor nem, ez az esetek nagy részében jól működik az ellenőrzőkben, de nem elhanyagolható részében hibáznak. Sok képzőt nem is vesznek fel a generatív rendszerbe, mert nem regulárisak, tehát nem nagyon lehet tudni, mikor alkalmazhatók. Másrészt, elemzők gyakran fogadnak (ajánlanak) nem használatos képzőformákat.(Naszódi, 2017)

Mint írtam, egyes esetekben formai módon eldönthető a kérdés. A műveltetés például *-ít* képző után mindig a *-tat, -tet* alakot várja. Ezt könnyen ellenőrizhetjük, ha lekérdezzük az összes *-ít* végű igét és megnézzük, mit szól az elemző/ellenőrző, ha a műveltetés különböző alakjait ráakasztjuk az igékre, illetve a nyelvérzékünk tiltakozik-e valamelyik alak ellen.

Még 1993-ban így jöttem rá, hogy az *[æ]*-re végződő melléknevek ritka kivétellel[2] megkaphatják az *-ít*, *-[uü]l* igeképzőt, ráadásul ilyenkor mindig elhagyjuk a szótővégi magánhangzót: *hülye* → *hülyül, hülyít*, *barna* → *barnul, barnít*... Ez a hangzókiesés a mellékneveknél szinte mindig jelentkezik, a főneveknél szinte soha, de algoritmikusan nem kezelik a helyesírás-ellenőrzők, hanem a képzett ige szerepel a szótőtárban, pedig a kategóriában reguláris szabály. Ezt az ismeretet fel is használtam a saját nyelvi adatbázisomban, illetve a helyesírás-ellenőrzőmben. Főneveknél viszont ritka kivétel, ha a szóvégi *[æ]* elnyelődik. Ha mégis, akkor felmerül a gyanú, hogy a szó valamikor melléknév lehetett.

A kérdés az, hogy mennyire ismerjük a toldalékainkat és szótöveinket, valamint mennyire lehet hasznos az ilyen egyszerű vizsgálat eredménye.

5. Teszteljünk!

Az algoritmus a következő: tapasztalatunk alapján legyenek gyanús toldalékolt szavak, toldalékok. Ezeket kigyűjtjük a szótárakból, leveszünk analitikus módon egy-két morfémát, majd generatív módon, figyelmbe véve a tőváltozást is, ráteszünk másikat, és megnézzük mi sül ki belőle. Az előállított szóformáról, ha nincsenek sokan, magunk is dönthetünk. Ha többen vannak, akkor bízzuk a meglevő elemzőkre, de teljesen sohase bízzunk a gépi döntésben.

[2] A forma, féle, fajta szavaink főnévként is funkcionálnak, illetve melléknévi névutóként kezelhetők. Más kivétel is található.

Példaként egy szintén 90-es évek eleji megoldott feladat.
Lemma: Ha az *-[aeoö](sz|j)t*, *-[aeoö]d* végű igék párban vannak (*fullaszt – fullad*), az első aktív (általában tárgyas), a második passzív (általában tárgyatlan) ige. Ráadásul az aktív és a passzív ige további *-ék* főnévképzőt, és további *-[eo]ny* melléknévképzőt kaphat (*halad-ék-ony*).

Egyszerű reguláris szűréssel 117 ilyen igepárt gyűjtöttem ki. Ezekre igaz volt az állítás, sőt, a kimaradtak, melyeknek csak egyik fele volt meg, többé-kevésbé állt a tárgyasságról való feltételezésem. Vizsgálva a további toldalékolást, ha nem is voltak szótári vagy legalább használatban levő szóalkotások, akkor is értelmes képzés volt: a *süllyeszt – süllyed* szópár alapján a *süllyeszték* még szakirodalmakban szerepel, de a *süllyedék*, *süllyedékeny* soha elő nem fordult, mégis jó magyar szó!

Azt is vizsgáltam, hogy ha a pár egyik fele szerepelt csak a nyelvi adataim közt, a másik alak milyen a nyelvérzékemnek. Nos van egy-két szójelöltem, ami teljesen magyar, még sincs a szótárakban. Az említett képzőket a jelenlegi szpellerek nem kezelik konstruktívan, mert nem minden igéhez társulhatnak. De az említett kategóriához igen.

Mellékesen egy kényes kérdésre választ kapunk, a műveltetést tisztázza az adott kategóriában: a párok *-d* végű alakjának az *-szt* végű alak a helyes műveltetett formája, és nem a *-t?[æ]t*.

6. Szótővadászat

Ha találtunk olyan képzőket, melyek valamire rátapadnak, akkor vizsgálhatjuk, a képzőtől megszabadított szó szótő-e. Ha megtesszük ezt a fent említett 117 esetre, akkor olyan eredményt kapunk, amivel nem tudunk sokat kezdeni. Azt tapasztaltam, hogy 9 esetben ige a szó eredendő töve, és 27 esetben névszó: főnév, melléknév vagy főnévi gyök. A maradék 80 is nyílván valamilyen ősi szavunk maradványa, de ezt egy vérbeli etimológus tudná csak igazolni, vagy ő sem. A talált 9+27=36 szótő viszont szerepel a szótárakban.

Érdemes a morfológiai rendszerbe felvett képzőkkel próbálkozni. Ha levágjuk egy szótári szóról, kapunk-e a várt szófajnak megfelelő már bejegyzett szó alakalternánsát. Ha igen – és a morfológiai rendszer ereje ettől nem csökken –, a képzett alakot elhagyhatjuk rendszerünkből, mert csak a redundanciát növeli.

Próbálkozhatunk a lexikonokban nem kodifikált toldalékok levágásával is. Ha már a fejezet címe **Szótővadászat**, kereshetjük az *-[áé]sz* képzős szavakat is, milyen szótőhöz tapadhatnak, és hogy viszonyulnak a további *-[æ]t* képzőhöz. Ha megkeressük, melyek az *-[áé]sz[æ]t* végű névszavaink, több mint 850-et találunk. Pontosabban ennél kevesebbet, mert ezek közt lesz olyan, amelyik csak formailag végződik úgy, mintha a két főnévi képzővel végződne a szó, de valójában más a morfológiai szerepe az elsőnek, például igei, csak a második főnévképző: *tenyészet*.

Ha a renitenseket kidobjuk, akkor a maradékban az első képző foglalkozást jelentő főnevet jelent, míg a második toldalék után vagy a tevékenységet, vagy

a tevékenység intézményét fejezi ki a szó. A képzők többnyire főnévhez kapcsolódnak: *rákász, fod(o)rász, jogász*, de néha igéhez: *szabász, szülész*. Itt is kereshetünk szógyököket, melyek többnyire főnévi jellegűek: *csábász*. Ha a második főnévképzővel nem szerepel a szó, óvatosabbak legyünk. Mert *csibészet* nincs, (bár lehetne).

Ebben a példában két dolgot tisztázhattunk. Az egyik, hogy nehéz megállapítani, hogy az *-[áé]sz* mihez kapcsolódhat, tehát jogos lehet ezen képzett szavak egyedi felvétele a szótárba. De ha felvesszük, a szótárban jelezhetjük, hogy miből származtatható, mert a képző szemantikai funkciója jól következtethető a képzésből. A másik tanulság, hogy ezek a szavak mind megkaphatják az általánosnak nem mondható *-[æ]t* főnévképzőt, és még a szemantika is származtatható algoritmikusan: az *-[áé]sz* olyan foglalkozást jelent, ami az alapszóval kapcsolatos. Fordítható úgy, hogy *a ... mestere* vagy *...-v[æ]l foglalkozó ember*: *fodrász = a (haj)fodor mestere*, *gyógyszerész = gyógyszerrel foglalkozó ember*. Az ezt követő főnévképző vagy magát a tevékenységet, vagy a tevékenység intézményét jelenti.[3]

7. Toldalékcsokrok

Az eddigi példákból is kiderült, hogy az egyes toldalékok nem mindig hatékonyak az elemzéshez. Ha hasonló szavakhoz jól illeszkedő toldalékcsoportot – csokrot – választunk, akkor biztosabb a vizsgálat eredménye. A korábbi példákban az *-ít* és az *-[uü]l* kétszálú csokor volt. A toldalékokból – mert még magyarban sincsenek olyan sokan – kevéselemű csokrokat képezhetünk. Nem így, ha a toldalék előtti lehetséges formákat keressük.

8. Szófürtök

Ha egy toldalékcsokorral a szótári bejegyzésekből tőgyanús alakot kapunk, akkor már érdemes megnézni, valóban találtunk-e valami rejtett szót. Ha különböző szótári tételek ily módon való csonkítása azonos szóhoz vezet, akkor ezek a korábban ismert szavak egy fürtöt alkotnak. A fenti példáim közt is akad ilyen.

Szófürtöt képezhetünk kodifikált (reguláris) és ritka toldalékok csokrának segítségével is. Ez utóbbinak igéink vizsgálatánál vettem nagy hasznát. Ha megnézzük az igei szótövek végeit, sokkal korlátozottabbak(Farkas és Naszódi, 1990), mint mellékneveinknél, főneveinknél. Szinte mindegyik képzett alak. Míg névszavakat a nyelvfejlődés során egyszerűen átvettünk velünk együtt élő népektől, igéink mindig valamilyen igeképző segítségével csapódtak nyelvkészletünkbe. Ez alól kivételek az ősi eredetű alapigéink: *esz(ik), alsz(ik), van, lesz, jön, megy*...

[3] Kivételek persze vannak. Nem minden főnevet követő *-ász -ész* képző ilyen: A *kolbász* szemantikailag nem foglalkozást jelent. Ha nem is találjuk egyik szótárunkban a *kolb, kolob, kalub* szavakat, van ilyen. A székelynek tartott, Benedek Elek által írásba foglalt *A kis gömböc* mese orosz változata: *Колобок*, aminek a jelentése azonos a kerek hentesáruval, akár a magyar gömböcé.

Ha pedig így van, kereshetjük, honnan származhatnak az igéink. Egyszerű mintaillesztéssel feltérképezhetjük a jellegzetes igevégződéseket – ezek alkalmasak lehetnek újabb képzők felfedezésére is – majd levágva, vizsgálhatjuk, kapunk-e szófürtöket. Ha jó a feltételezés, akkor a szófürtöket a képzőhalmazok határozzák meg. Pár igevégződést kiválasztva vizsgálhatom az így kapott fürtöt.

8.1. Alaki szófürtök alkalmazásai

Formai hasonlóságon alapuló szófürtöket gyakran használnak praktikus céllal. A sémi nyelvek szótárainak tételei szófürtökön alapulnak: az inflexiós szabályok miatt a nálunk szokásos ábécébe rendezés miatt szócsaládok kerülhetnek távol egymástól, pedig azonos a gyökük, szótövük. A szó szerkezetéből kihámozható szógyök ábécébe rendezése az elsődleges sorba rakási elv, aminek karakterei nem feltétlen a szó elején találhatók. Ez persze egy nehezebb mintaillesztés, mint ami nálunk használható, nem szóeleji vagy szóvégi hasonlóság keresése.

A Microsoft keresőrendszerében nyelvi támogatásként egy klaszterező rendszer volt. Főként formai ismérvek alapján társított szóalakokat. (Lehet, hogy ma is használják ezt a módszert.) A klaszterezés nyelvenként változott. Így a ragozott, képzett alakok nagy valószínűséggel egy fürtbe kerültek. A magyar nyelvre sehogy sem működött a módszer, de latin nyelveknél kiváló volt – a keresők ebből eredő hibáját az ember felülbírálhatta.

Több kísérlet volt a világon, hogy formai ismérvek alapján építsék ki egy nyelv ragozási rendszerét. Egyik sem volt tökéletes, de hatékony.(Wicentowski, 2004) Az egyik tanítványom például nagy korpusz alapján hozott össze olasz ragozási szótárat – osztályozta a szavakat: a szóvégek alapján előállított toldalékcsokrokat, majd keresett ezekhez tartozó szófürtöket. Így szerkesztette meg a szavak teljes ragozási paradigmarendszerét. A magyarhoz hasonló összetettségű szószerkezettel rendelkező nyelv (török) morfológiájának felépítésére is sikeresen alkalmaztak hasonló módszert.(Oflazer és Nirenburg, 1999), de a török fonológiája egyszerűbb, mint a magyaré.

Ha már létezik jó ragozási beosztás, korpuszból kinyert ismeretlen szóformákat lehet besorolni(Novák és mtsai, 2003), ha sikerül jól összehozni a szófürtöket... A magyar szó szerkezete és fonológiája elég összetett ahhoz, hogy automatikus rendszerbe ne bízzunk, de arra megfelel a módszer, hogy jó tanácsokat kapjunk.

8.2. Példa

Hogy bemutassam az eljárás hatékonyságát, a végződéshalmazt most önkényesen választom ki – az lesz a feltételezésem, hogy a következők igeképzők. A kiválasztott igevégek: *-[œ]n*, *-[œ]nt*, *-[aeoö]g*. Egyik sem szerepel morfológiai algoritmusokban.

Az így végződő találatokból kidobva az egy szótagúakat – ezek nem lehetnek képzettek, hisz a képző is egy szótag – első vizsgálatra feltűnik pár tulajdonság.

– A *-g* végű ige kivétel nélkül ismétlődő passzív (szenvedő), vagy legalábbis hosszan tartó, tárgyatlan.

– Az -*n* végű ige rövid idejű, inkább passzív (általában tárgyatlan).
– Az -*nt* végű ige mindegyike rövid idejű aktív (általában tárgyas).
– Ha a fürt teljes (mindhárom forma szerepel az eddigi szótárban), az igék alapjelentése azonos, csak paraméterei, vonzatai cserélődnek, tehát az egyik tárgya a másiknak alanya, stb. . .
– Ha megleljük a szótövet, akkor világos, ebből képeztetnek az igék.

De nincs új a nap alatt, hisz a nyelvészeket régóta izgatja a kérdés, a képzőket régóta igyekszenek feltérképezni.(Ihász, 1846) Így megállapították, hogy hangutánzó szavainkból, főleg az egytagúakból így képezhetünk igéket.

Kérdezzük le a nyelvi adatbázist, igaz-e a feltételezés, illetve mely szavaink lehetnek hangutánzók: keressük a dupla mássalhangzóra végződő főneveket, esetleg indulatszavakat, mondatszókat, határozószókat, és vessük össze azzal a fürttel, melyek elemei a korábbi három képzőformához tartoznak. Én 47 olyan szót találtam, amely a csokorhoz részben passzol, vagyis van olyan ige a fürtben, ami az újonnan kigyűjtött adott hangutánzógyanús szóból származik. Ezek többsége teljes toldalékcsokorhoz passzol. Sőt, ami tényleg hangutánzó, azoknak mind teljes a csokra.

Az alábbi táblázatomba belevettem -*j*-re végződő zajra, hangra vonatkozó szavakat is (felkiáltójel). Utána némi szubjektív szűréssel felvettem szótárakban nem szereplő alakokat is, hogy teljesebbé tegyem a fürtöket. A -*g* vég helyett megengedtem az -*ng* végződést, esetenként más tőváltozatot. Bár gyakran az -*/æ/nt* helyett szebben hangozna az -*int* igeképző, ezeket nem jeleztem. Ezek általában más csokorhoz jobban kapcsolódnak, melyeket most nem vizsgáltam. Pár szónál kérdőjellel jelöltem a számomra bizonytalan alakokat, még ha kodifikált szótári tételek is voltak.

Nem minden szó hangutánzó, de ha teljes volt a csokor, benne hagytam, mert a jelentésen kívüli egyéb nyelvtani tulajdonságok megfeleltek a feltételezésemnek. Sok olyan hiányos csokrot is benne hagytam a listába, melyek hanghatást fednek. Ha teljes volt a paradigma, alapszavát kiemeltem. Többségben vannak! Ahol az alapszó hiányzik, ott szógyök keresendő.[4]

alapszó	passzív	aktív	ismétlő		alapszó	passzív	aktív	ismétlő
berr			berreg			biccen	biccent	biceg
	billen	billent	billeg		böff	böffen	böffent	böfög
brekk	brekken		brekeg		brumm			brummog
buff	buffan	buffant	buf(f)og		**buggy**	buggyan	buggyant	bugyog
büff	büffen	büffent	büfög		camm?			cammog
cin			cincog		cö			cöcög
			csacsog		cupp	cuppan	cuppant	cuppog
csatt	csattan	csattant	csattog		**csepp**	cseppen	cseppent	csep(er)eg
cserr	cserren		cserreg		**csett**	csetten	csettent	csetteg
csevej!			cseveg					csicsereg
	csillan	csillant	csillog		csipp	csippen?	csippent	csipeg
csipp	csippen?	csippent	csipog		**csirr**	csirren	csirrent	csir(r)eg
csissz	csisszen		csiszeg		**csitt**	csitten	csittent	csitteg
			csivog		**csobb**	csobban	csobbant	csobog
csorr?	csorran?	csorrant?	csorog		**csossz**	csosszan		csoszog
csöpp	csöppen	csöppent	csöp(ör)ög		**csörr**	csörren	csörent	csörög
alapszó	passzív	aktív	ismétlő		alapszó	passzív	aktív	ismétlő

[4] szógyök olyan morféma, amely többnyire csak szóösszetétel első tagjaként használható, vagy úgy sem, de nyelvészeti eszközök igazolják önálló jelentéssel bíró létezését.

alapszó	passzív	aktív	ismétlő
csurr	csurran	csurrant	csurog
	rezzen	rezzent	rezeg
dirr	dirren	dirrent	dirreg
	döbben	döbbent	döbög
			dörmög
			dunnyog
	duzzan		duzzog
			dünnyög
fitty			fityeg
	fortyan	fortyant	fortyog
fütty	füttyen	füttyent	füttyög
			habog
háp			hápog
	herren		herreg
	hortyan		hortyog
huh			huhog
hüm(m)			hümmög
hess	hessen	hessent	
hipp	hippen		hipeg
hupp	huppan		huppog
kacaj!			kacag
katt	kattan	kattant	kattog
ketty	kettyen	kettyent	ketyeg
kocc	koccan	koccant	kocog
kukk	kukkan	kukkant	
kotty	kottyan	kottyant	kotyog
			krákog
kurr?	kurjan	kurjant	kurrog
	libben	libbent	libeg
loccs	loccsan	loccsant	locsog
lötty	löttyen	löttyent	lötyög
mekk			mekeg
moraj!	morran	morrant?	mor(m)og
	mozzan		mozog
nyaff	nyaffan		nyafog
nyekk	nyekken	nyekkent	nyek(er)eg
nyiff	nyiffan	nyiffant	nyifog
nyikk?	nyikkan		nyik(or)og
	percen		perceg
petty	pettyen	pettyent	petyeg?
	pihen		piheg
	pislan?	pislant	pislog
pitty	pittyen	pittyent	pityereg
			porcog
potty	pottyan	pottyant	potyog
	pörren	pörrent	pörög
	prüsszen	prüsszent	prüsszög
			pusmog
	rebben	rebbent	rebeg
	rekken	rekkent?	rekeg
	rezzen	rezzent	rezeg
	rohan		rohog
			roszog
röf(f)	röffen	röffent?	röfög
rötty	röttyen	röttyent	rötyög
			selypeg
	sercen	sercent	serceg
			sipeg
slatty			slattyog
			sunnyog/sunyorog
	sussan?		sus(m)og
			suttog
sutty	suttyan		suty(or)og

alapszó	passzív	aktív	ismétlő
csüccs	csüccsen	csüccsent?	
robaj!	robban	robbant	robog
dobb?	dobban	dobbant	dobog
döcc	döccen		döcög
dörej!	dörren	dörrent	dörög
durr	durran	durrant	durrog
			düb(ör)ög
			fecseg
forr	forran	forrant	forrong
fröccs	fröccsen	fröccsent	fröcsög
gá			gágog
	harsan		harsog
			hebeg
	hersen		herseg
	hörren		hörög
hurr?			hurrog
hüpp	hüppen		hüppög
hepp			hepeg
hopp			
huss	hussan		
	kaffan	kaffant?	kaffog
kár			károg
kipp	kippen	kippent	kipeg
kopp	koppan	koppant	kopog
	korran?		korog
?köhej!	köhhen	köhhent	köhög
			kunc(or)og
	lebben	lebbent	lebeg
lob	lobban	lobbant	lobog
lotty	lottyan	lottyant	lotyog
makk			makog
mocc	moccan	moccant	mocorog
	mottyan		moty(or)og
mukk	mukkan	mukkant	
nyau			nyávog
			nyervog
			nyihog
patt	pattan	pattant?	pattog
			pereg
piff	piffen	piffent	pifeg
	pillan?	pillant	pillog
pissz	pisszen	pisszent	pi(s)szeg
			pizseg
			poszog
pöff	pöffen	pöffent	pöfög
pötty	pöttyen	pöttyent	pötyög
puff	puffan	puffant	pufog
püff	püffen	püffent	püfög
reccs	reccsen	reccsent?	recseg
	retten	rettent	retteg
robaj!	robban	robbant	robog
ropp	roppan	roppant	ropog
rotty	rottyan	rottyant	rotyog
röhej!			röhög
sáp			sápog
	seppen	seppent	seppeg
	settyen		settyeg
			sistereg
	suhan		suhog
surr?	surran	surrant	surrog
			sustorog
			sutyorog
	szeppen		szepeg

alapszó	passzív	aktív	ismétlő
		szippant	szipog
	szortyan		szortyog
			szörtyög
szotty	szottyan		
toccs	toccsan	toccsant?	tocsog
totty	tottyan		totyog
trapp			trappog
			vacog
			varcog
			vernyog
			vigyorog
			vijjog
			vinnyog
zizz?	zizzen	zizzent	zizeg
zökk	zökken	zökkent	zökög
zötty	zöttyen	zöttyent	zötyög
	zuhan		zuhog
zupp	zuppan?		
zsibaj!			zsibo(n)g
	zsizzsen		zsizseg

alapszó	passzív	aktív	ismétlő
szissz	szisszen	szisszent	szisz(er)eg
			szörcsög
szusz	szusszan	szusszant	szuszog
tipp	tippen	tippent?	tipeg
topp	toppan	toppant	top(or)og
töf(f)			töfög
	tüsszen	tüsszent	tüsszög
vakk?	vakkan	vakkant	vakog
	vartyan		vartyog
			vicsorog
			vihog
	villan	villant	villog
zaj			zajo(n)g
zok			zokog
zörej!	zörren	zörent	zörög
			zub(or)og
zümm			zümmög
			zsenyeg
zsivaj!			zsivo(n)g

9. Egyéb szótani rejtelmek

A példáim főként arra szolgáltak, hogy algoritmizálhassunk eddig figyelembe nem vett képzőket. Nem csak erre jó. Most felsorolom, én hol vetném be a magyar nyelv esetén:

- Összetételek keresése: szótári tételek akár szó eleji, de inkább szó végi egyezése alapján képeznék szófürtöket, melyeknek szófürtképző alakjai szintén szótári tételek.
- Új képzők keresése: ha alaki azonosság alapján találunk olyan fürtmeghatározó szóvégcsokrot, amelyhez nagy szófürt tartozik, feltehetően a fürtképző karakterláncok toldalékokból állnak.
- Szótőkeresés: ha az előző módszer toldalékcsokraihoz tarozó szójelölt többsége valóban helyes, akkor a szótőként korábban nem értelmezett fürtelemek, (az a karakterlánc, melyre a toldalék elemei csatlakoznak) vizsgálható, jó-e szótőnek, esetleg gyöknek.
- A fenti két módszer iteratív alkalmazásával akár ismeretlen nyelv morfológiáját is felfedhetjük, vagyis kialakíthatunk teljes toldalékcsokrokat, és megállapíthatjuk az ezekhez tartozó szótőfürtüket, vagyis az egy toldalékolási paradigmához tartozó szócsoportokat.[5]
- Szógyökkeresés: egy időben gyűjtöttem az élő szógyököket. Élő a szógyök, ha önállóan nem, legfeljebb ragozva, képezve, egyébként összetételben fordul elő (*gyógyot, gyógykezelés, gyógyul*). Halott, ha csak képzett alakban van jelen (*segít, segély*).

Ezekkel a módszerekkel sok egyéb nyelvi finomság is napfényre kerülhet.

Had emeljem ki: az esetek többségében nem szövegkorpuszokat használok, hanem nyelvi, szótárjellegű korpuszokat. Ezekből nyerem az információkat, hogy visszacsatolva javítsam a minőséget. Ez szemben áll a mai gyakorlattal, mert

[5] http://wordlist.aspell.net/agid-readme/

szószinten azt tekintik a nyelvészek bizonyítéknak, ha valami elő is fordul. Ezzel szemben a valóság az, hogy helytelen dolgok is előfordulnak, de a nyelv része az is, ami még soha le nem lett jegyezve, de akár lehetne is. Mondatszinten ez természetes, hisz a helyes mondatok variációja akkora, hogy számba sem lehet venni. Az aglutináló nyelveknél szószinten is igaz az az elv, hogy nem csak az van, ami elfordult valamikor, tehát nyoma van korpuszban. Nem csak azok a szerkezetek léteznek, melyeket eddig felfedtünk. Papp Ferencet kéne idéznem, de nem tudom mikor és hol mondta. A pontos szövegére sem emlékszem, de a lényege a következő: *Mi tartozik a magyar nyelvhez? Az a mi valamikor elhangzott vagy leírták? Vagy az is. ami el fog hangzani? Vagy amit valaki akár el is mondhat?* A lényeg ebből a szempontból az, hogy szótanunk kellően összetett, hogy olyan eszközöket használjunk, amelyeket sok nyelvben csak mondattannál alkalmaznak.

A 90-es évek elején vizsgáltam, hogy a nyelvészek a különböző tőváltozatokat hogyan osztályozzák. Toldalékolásnál, ragozásnál fontos ismérv, hogy mely toldalék melyik alakja melyik tőalternánshoz kapcsolódik. Ebből a szempontból azonosnak bizonyult a szóbelseji magánhangzó-rövidülés (*madár, madarat*) és a belső hangzókiesés (*szatyor, szatyrot*), sőt, az úgynevezett mássalhangzó-átvetés (*teher, terhet*) teljesen egybeeső paradigma.[6] A ragozás algoritmusa teljesen egységesen fogalmazható meg ezekben a nyelvészek által különválasztott kategóriákban. Némi fenntartással a *-v* betoldásával járó tőváltozás is ide sorolható (*daru, darvat; tó, tavat*), de itt vannak apró különbségek.

Akár meglevő morfológiai szótárak fésülésére is alkalmazhatjuk az eljárást.[7] Még sok ötletem van. Utolsónak a mifantológusoknak, történeti nyelvészeknek ajánlom figyelmébe, hogy használják bátran az egyszerű szűrést. Ily módon esett le nálam a tantusz, hogy a *tavasz* szavunk minden bizonnyal a *tó* képzett alakja. Nyelvészektől (is) hallottam, hogy a szóhasználat következtében főnév melléknévvé, melléknév főnévvé változhat az idők során. A legszebb példát Prószéky Gábortól: a *róka* állítólag eredetileg melléknév volt, míg a *ravasz* főnévként kezdte nyelvünkben az életét. És mindkét szó ugyanonnan származik. Az egyik a *-ka* kicsinyítő képzőt kapta, a másikat az *-asz* képző díszíti. Ezek szerint mindkettő töve a *ró* főnév, csak az egyik tőváltozás nélküli, a másik egy gyakori *-v* kötéssel

[6] Elekfi László ragozási szótárában az 5-ös, 7-es és a 9-es névszói fő osztályok egy kalap alá vonhatók. Ezek a Hunspell adatbázisában is külön osztályt képeznek, pedig a formalizmus lehetővé tenné az egységes kezelést. Ha a hosszú magánhangzót kettőzött karakterrel jeleznénk, mint ahogy a finnek teszik, triviálissá válna, hogy nincs különbség a hangzókiesés és a hangzórövidülés között.

[7] A magyar helyesírás-ellenőrzők hőskorából származik az a történet, hogy – mivel az értelmező kéziszótár sem tartalmazott mindent – gyakran bővítettük a tőtárat. Az egyik ilyen volt a *csevej*. Pontosabban, nem találtuk, hogyan kell írni, és a döntés a *csevely* alak lett. Pár hónapig ebben hittek a fejlesztők, és a felhasználókat is erre irányította a program. Az irodalomban, mint kiderült, mindkét forma előfordult, mert első látásra nem triviális. Emiatt, ha akkor lettek volna nagy digitális korpuszok, akkor sem tudtunk volna dönteni. Most viszont, a (8.2.) fejezet táblázatába foglalt szavak mutatják: ha az *-[æ]j* véget levágjuk, és a megmaradt részre az *-[aeoö]g* toldalékot tesszük, akkor jó igét kapunk. Ez az *-ly* végű főnevekre nem áll.

jön létre. Ilyen főnevünk nincs, de a zürjén *rutjs*, a cseremisz *revezs*, a mordvin *rives*, a finn *repo rókát* jelent(Tótfalusi, 2013).

10. Összefoglalás

A számítógépes nyelvészetben gyakran egyszerű alaki tulajdonságok rejtenek fontos nyelvi információkat. A morfológia esetén, már régóta használnak egyszerű formai szűréseken alapuló osztályozásokat, elemzéseket. Ennek egyik kiterjesztése, ha nem mintákkal, hanem mintafürtökkel szűrünk. Ezzel a kapott szófürtökről pontosabban dönthetünk.

A módszer általánosan alkalmazható, és sok helyen lehet hasznos. Én csak pár példát mutattam meg, de bátorítom a nyelvészeket, éljenek a számítógépek lehetőségeivel ötletesen, mert nagyon megéri.

Hivatkozások

Elekfi, L.: Magyar ragozási szótár. MTA Nyelvtudományi Intézete (1994)

Farkas, E., Naszódi, M.: A toldalékok 32 fonológiai osztálya. In: Magyar nyelvű mondatok elemzése természetes nyelvű interfész céljából. p. 44. MTA SzTAKI (1990), http://www.cs.bme.hu/~naso/langeng/manyel.pdf

Ihász, r.: Igeképzők. In: Magyar nyelvtan. p. 213-220 (1846)

Koskenniemi, K.: Two-Level Morphology: A General Computation Model for Word-Form Recognition and Production. No. 11 in PUBLICATIONS, University of Helsinki, Department of Linguistics (1983), http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf

MTA, N.I.: Magyar értelmező kéziszótár. Akadémiai Kiadó (1972)

Naszódi, M.: A magyar helyesírás-ellenőrzők mai állása. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia. p. 347-354. Szegedi Tudományegyetem (2017), http://www.cs.bme.hu/~naso/langeng/SpellsSate20016.pdf

Novák, A., Nagy, V., Oravecz, C.: Magyar ismeretlenszó-elemző program fejlesztése. In: Magyar Számítógépes Nyelvészeti Konferencia. p. 45-57. Szegedi Tudományegyetem (2003)

Oflazer, K., Nirenburg, S.: Practical bootstrapping of morphological analyzers. In: Conference on Natural Language Learning (1999)

Papp, F.: A magyar nyelv szóvégmutató szótára. Akadémiai Kiadó (1969)

Proszéki, G., Kornai, A.: Papp Ferenc és az újrafelhasznált Szóvégmutató szótár (2017), https://itf.njszt.hu/objektum/papp-ferenc-es-az-ujrafelhasznalt-szovegmutato-szotar

Tótfalusi, I.: Magyar etimológiai nagyszótár. Arcanum (2013), http://www.szokincshalo.hu/szotar/

Wicentowski, R.: Multilingual Noise-Robust Supervised Morphological Analysis using the WordFrame Model. In: ACL Special Interest Group on Computational Phonology (SIGPHON) (2004), https://www.aclweb.org/anthology/W04-0109.pdf

A Note on Lost Morphemes
The Benefits on Surface Similarities

Mátyás Naszódi

MorphoLogic KFT.
naszodim@morphologic.hu

Abstract. The current morphological analyzers have been designed pragmatically for practical purposes. Their goal is to cover the word forms in Hungarian texts with relatively little effort and with as few mistakes as possible. Once the goal has been achieved, regular case affixes, marks, and verbal conjugation endings are well described in a formal way, but most derivative affixes and rare case suffixes are treated individually as exceptions.

In my research, I found that there are far fewer exceptional word forms in Hungarian. By clustering word forms by their endings, new relationships, new roots, new morphemes can be discovered that are missing from earlier databases. By clustering word forms by their endings, new relationships among roots, morphemes can be discovered that are missing from earlier databases. One can simplify morphological descriptions without limiting their power. Even a complete morphological description of an unknown language can be generated based on a large corpus solely. Moreover, if not only similarities of endings, but clusters of ending patterns are used to group word forms, then many hidden word roots and suffixes can be discovered that have been forgotten altogether, or mentioned only by descriptive linguists.

As a result of the method, semantic dependences might be discovered, and linguistic collections, databases made for practical purposes can be corrected, improved as well.

Keywords: vocabulary, morphology, lexicography, spelling

Beszédtechnológia II.

Az akusztikus szózsák eljárás korpuszfüggetlenségének vizsgálata

Vetráb Mercedes[1], Gosztolya Gábor[1,2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
Szeged, Árpád tér 2.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
{ vetrabm, ggabor } @ inf.u-szeged.hu

Kivonat Cikkünkben egy jellemzőreprezentációs módszer, az akusztikus szózsák (Bag of Audio Words, BoAW) metódus szélesebb körű használhatóságát elemezzük. A BoAW eljárás lehetővé teszi a változó hosszúságú hangminták fix méretű jellemzővektorokként való kezelését. Ezáltal a különböző hangadatbázisok kezelhetővé és taníthatóvá válnak a hagyományos tanulóalgoritmusokkal is. A BoAW eljárás kezdeti lépésében klaszterközéppontokat (ún. kódszavakat) határozunk meg a keretszintű jellemzővektorok fölött valamilyen felügyelet nélküli módszerrel (pl. k-means klaszterezéssel, vagy akár csak véletlenszerű kiválasztással). Ezt a lépést hagyományosan az adott akusztikus adatbázis tanító halmazán szokás elvégezni. Ez azonban amellett, hogy minden adatbázison új kódszavak kiválasztását teszi szükségessé, így megnyújtva a jellemzőreprezentációk előállításának idejét, akár túllillesztést is okozhat. Jelen tanulmányunkban megvizsgáljuk, hogy mennyire korpuszfüggő az előálló kódszóhalmaz. Kísérleteinkben egy magyar nyelvű érzelemadatbázison mérünk osztályozási eredményeket, miközben a kódszavak kiválasztása vagy egy német nyelvű érzelemadatbázison, vagy egy magyar nyelvű, általános beszédadatbázison történik. Eredményeink szerint mindkét új típusú megközelítéssel elérhető, a korábban említett hagyományos megközelítéssel elérhető osztályozási pontosság, ami megkönnyítheti a BoAW eljárás gyakorlati alkalmazását.
Kulcsszavak: érzelemfelismerés, hangfeldolgozás, akusztikus szózsák reprezentáció

1. Bevezetés

Az emberi beszéd a közlendő szövegen túl egyéb információkat is magában hordoz. Árulkodhat akár a beszélő fizikai vagy emocionális állapotáról is. Napjainkban az automatikus érzelemdetektálás egy folyamatosan fejlődő ágazat. Technikai alkalmazási köre igen széles skálán mozog. Többek közt hasznos az ember-gép interakciók során (az ember kommunikációjának monitorozására) (James és mtsai, 2018), dialógusrendszereknél (Burkhardt és mtsai, 2009), az egészségi állapot

felméréseknél (Hossain és Muhammad, 2015; Norhafizah és mtsai, 2017), valamint a call-centerekben (Vidrascu és Devillers, 2005). Ezen terület fejlesztése és kutatási eredményeinek előrelépése akár hétköznapi rendszereink jelentős fejlődését is eredményezheti.

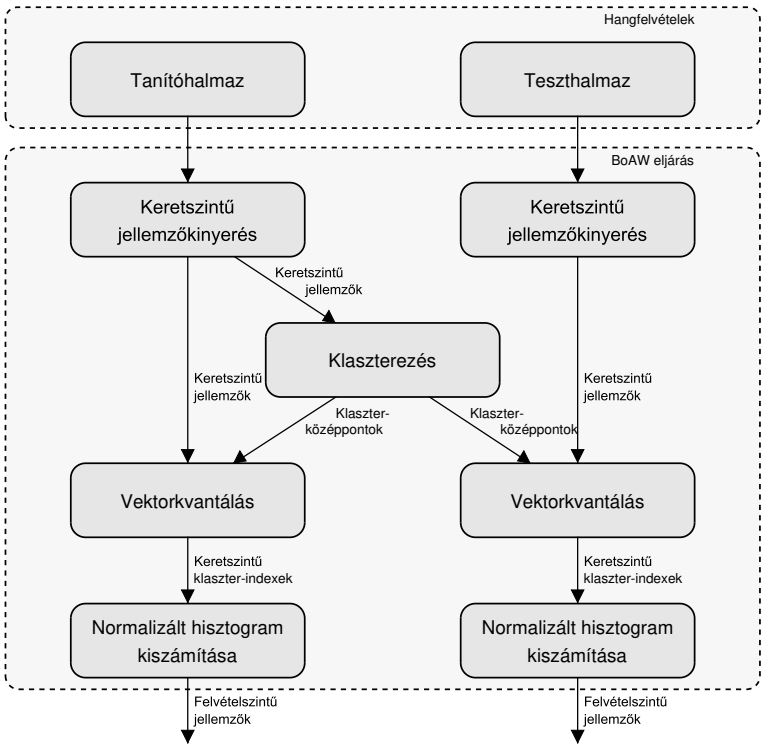
A terület kutatásának kezdete óta több módszert is kidolgoztak arra nézve, hogy a hangfelvételekből milyen módon érdemes jellemzőket kivonni, valamint arra, hogy melyek azok a tanulóalgoritmusok, amik a legoptimálisabb és legeffektívebb eredményeket szolgáltatják egy-egy mintahalmazon. Ezen cikk alapját egy korábbi tanulmányunk adta (Vetráb és Gosztolya, 2019), ahol az akusztikus szózsák (Bag-of-Audio-Words, BoAW (Pancoast és Akbacak, 2012)) technikát és annak sikerességét vizsgáltuk. Pozitív eredményeink alapján újabb kérdések merültek fel, melyeket az alábbi tesztek során igyekeztünk megválaszolni. Korábban a BoAW technika egyik legnagyobb hátrányának a jellemzők előállításához használatos kódhalmaz (codebook) előállításának hosszú idejét tapasztaltuk, melyre megoldást jelentene, ha lehetőség nyílna egy már előre meghatározott codebookot egy másik adatbázisban is felhasználni.

Jelen tanulmányunkban azt vizsgáljuk, hogy más adatbázisból előállított codebookkal lehet-e hasonló vagy jobb eredményeket elérni, mint az eredeti adatbázisból előállított használatával. Mivel eredményeink alapján pozitív választ kaptunk, így megfogalmaztuk következő lényeges felvetésünket, miszerint ha hasonló feladatra készült, de különböző akusztikájú adatbázissal teljesítménybeli javulást értünk el, akkor vajon lehet-e hasonlóan jó eredményeket elérni, egy eltérő feladatra készült adatbázis codebook-jának használatával.

2. Az akusztikus szózsák eljárás

Az általunk használt *akusztikus szózsák* technika, azaz a *bag of audio words* (vagy BoAW) hasonló a szövegfeldolgozásban ismert *bag of words* és a képfeldolgozásban alkalmazott *bag of visual words* (BoVW) módszerekhez. Az 1. ábrán látható, hogy a BOAW módszer egyes fázisaiban végrehajtott műveleteket mind a tanító, mind a teszt halmazon elvégezzük. Első lépésben a tanítóhalmaz hangfelvételeiből kinyerjük az előre meghatározott jellemzőket, melyekből minden kerethez egy-egy jellemzővektor áll elő (keretszintű jellemzők). Ezután a jellemzővektorokból klaszterezés (felügyelet nélküli módszer) segítségével elkészül a kódszavak halmaza (kódhalmaz, *codebook*). A folyamat során megadott számú csoportot hozunk létre, ahol a klaszterek középpontjai lesznek a kódszavak (codewords). A csoportok számát nevezzük a codebook méretének; ez a szózsák eljárás egyik paramétere is.

A következő lépés a vektorkvantálás, mely során az egyes felvételekhez tartozó keretszintű jellemzővektorokat kvantáljuk az előző lépésben generált kódszavaktól vett minimális euklideszi távolságuk alapján. Az eredeti jellemzővektorok helyettesítésre kerülnek a hozzájuk legközelebb lévő kódszó indexével. Végül egy hisztogramot készítünk a kódszavak és hozzájuk sorolt vektorok gyakoriságából. Ebből adódóan a hisztogram mérete megegyezik a codebook méretével, és függetlenné válik az adott hangfelvétel hosszától. Az így előállított vektorhalmaz lesz



1. ábra: Az akusztikus szózsák eljárás működési módja.

a „*bag of audio words*” reprezentáció, ami majd a tanító algoritmusunk inputjául szolgál.

Az 1. ábrán látható, hogy a teszthalmaz esetében a klaszterezés lépése kimarad. Ezt azért tehetjük meg, mert az BoaW eljárás lehetőséget nyújt arra, hogy a tanítóhalmazhoz alkalmazott paraméterbeállításokat és az elkészült codebookot lementsük, majd később ezt használjuk fel (akár más adatbázisokból származó) további hangfelvételek akusztikus szózsák jellemzőreprezentációjának előállításához. Hiszen attól függetlenül, hogy az új mintahalmaz felvételeit nem használtuk fel a klaszterezés során, a keretszintű jellemzővektoraikat még besorolhatjuk az egyes klaszterekbe a kódszavaktól vett távolságuk alapján.

3. Kísérleti körülmények

A következőkben bemutatjuk az elvégzett kísérletek technikai körülményeit: az alkalmazott adatbázisokat és azok szerepét, az osztályozási eljárást és paramétereit, a kiértékelésre használt metrikát, valamint a keretszintű jellemzőkészletet.

3.1. Adatbázisok

A kutatás során minden esetben a magyar érzelemadatbázis tanító és teszthalmazának akusztikus szózsák jellemzőreprezentációján végezzük az osztályozást és kiértékelést. A két további adatbázis a kódszavak előállításában kapott szerepet.

Magyar érzelemadatbázis Az adatbázis 97 magyar anyanyelvű és magyarul beszélő személy hangját tartalmazza (Sztahó és mtsai, 2011). A beszédek televíziós műsorok során lettek felvéve. A szegmensek túlnyomó része érzelmekben gazdag, folyamatos, spontán beszédből lett kivágva. Kisebb részük improvizációs szórakoztató műsorból jön. Az adatbázis összesen 1111 mondatot tartalmaz, melyek egy 831 elemű tanító és 280 elemű teszt halmazra lettek osztva. A címkék négyféle érzelmet definiálnak a beszédekben: Harag, Öröm, Szomorúság és Semleges hangulat. A tanító adatbázis mintáinak eloszlása $\approx$ 57% semleges, $\approx$ 6% bánat, $\approx$ 9% öröm és $\approx$ 27% harag. A teszt adatbázis mintáinak eloszlása: $\approx$ 62% semleges, $\approx$ 4% bánat $\approx$ 7% öröm és $\approx$ 27% harag. A tanító adatbázis felvételeinek teljes hossza nagyjából 20 perc, míg a tesztfelvételeké nagyjából 7 perc.

Német érzelemadatbázis Az adatbázis 10 német anyanyelvű és németül beszélő személy hangját tartalmazza. A felvételeket 25 és 35 év közötti színészekkel készítették el. Minden alany 10 különböző mondatot mondott fel, mindegyiket több különböző érzelmi töltettel. A címkék hétféle érzelmet definiálnak a beszédekben: semleges, düh, unalom, undor, félelem, boldogság, szomorúság. Az adatbázis felvételeinek teljes hossza körülbelül 25 perc (Burkhardt és mtsai, 2005).

Magyar híradófelvételek adatbázisa Az adatbázis 8 különböző magyar nyelvű TV-s híradóműsor felvételit tartalmazza. Összesen 28 órányi felvételből áll. A felvételeken hivatásos televíziós műsorvezetők hallhatók, így érzelemdetektálás szempontjából, a hírek felolvasása érzelemmentesnek, azaz semlegesnek tekinthető. Az adatbázisban szereplő híradók felvételeit a felhasználás során először összekevertük, majd több különböző hosszúságú halmazra bontottuk: első 1 óra felvételei, első 2 óra felvételei, első 5 óra felvételei, első 10 óra felvételei. (Grósz és Tóth, 2013)

3.2. Osztályozás

Az osztályozást SVM-ek (Support Vector Machines (Schölkopf és mtsai, 2001)) használatával végeztük, lineáris kernellel, a libSVM implementációt használva (Chang és Lin, 2011). Az algoritmus komplexitás (complexity, C) paraméterét minden futtatás esetén többféle beállítással teszteltük. A lehetséges konfigurációk az alábbi 10 hatványok voltak: $0,00001$; $0,0001$; $0,001$; $0,01$; $0,1$ és 1. Egy adott modell „jóságának” mérésére az UAR metrikát (Unweighted Average

Recall: az adott osztályra helyesen osztályozott példák száma osztva az adott osztályba tartozó példák számával) alkalmaztuk.

A tanító halmazzal való munka során az osztályozó tanulását és kiértékelését 10-szeres keresztvalidálással (10-fold cross-validation, CV) hajtottuk végre. Tehát az aktuális mintahalmazt 10 közel egyenlő részre osztottuk úgy, hogy az azonos beszélőhöz tartozó minták, azonos fold-ba kerüljenek. Ezután minden lehetséges 9 tanító – 1 tesztelő halmaz kombinációra tanítottunk és kiértékeltünk egy SVM modellt. Később, a kereszvalidálás során a tanító halmazra kapott értékek alapján választottuk ki, hogy a teszteket milyen beállításokkal futtassuk.

A teszthalmazra adott predikcióinkat a teljes tanítóhalmazon tanított SVM modellek szolgáltatták.

3.3. Keretszintű jellemzőkészlet

Az akusztikus keretszintű jellemzők megválasztásának alapját a 2013-as INTERSPEECH Számítógépes Paralingvisztikai Versenyen kiadott cikk adta (Schuller és mtsai, 2013). Az ott publikált jellemzőkészlet 65 keretszintű jellemzőt, azaz LLD-t (low level descriptor) tartalmazott (4 darab energián alapuló LLD; 55 spektrális LLD; 6 hangosságon alapuló LLD), valamint ezek első fokú deriváltjait. Kutatásunk során ezen jellemzőket az openSMILE nevű program segítségével számoltuk le. A hangosság alapú leírókat 60 ms-os kerettel (Gaussian ablakfüggvény) és $0,4$ értékű szigmával, a másik két csoportot pedig 25 ms-os kerettel (Hamming ablakfüggvény) számítottuk. A Hamming-ablakokat a megszokott módon, átfedéssel, 10 ms-os eltolással helyeztük el.

3.4. Az akusztikus szózsák eljárás tesztelt paraméterei

Az akusztikus szózsák reprezentáció számítását megvalósító openXBOW (Schmitt és Schuller, 2017) program, a BoAW eljárás több, szabadon állítható paraméterének beállítási lehetőségével rendelkezik. Kísérleteink során teszteltük az adatok előfeldolgozásának, a codebook méretének valamint a kvantáláskor figyelembe vett legközelebbi szomszédok számának hatásait.

Egyik, említett beállítási lehetőségünkkel az elkészítendő codebook méretét adhatjuk meg. Itt határozzuk meg, hogy a klaszterezés során, hány klasztert állítsunk elő, tehát hány kódszót határozzunk meg.

Egy másik szabályozható komponens a hisztogram előállításának módja. Korábbi cikkünkben (Vetráb és Gosztolya, 2019) született megerősítő eredményeink alapján egyértelmű volt számunkra, hogy minden kerethez a legközelebbi klaszter helyett a legközelebbi a db. klasztert rendeljük hozzá, mivel így azonos méretű jellemzővektor mellett pontosabban írhatjuk le az adott felvételt.

Az előző módosításon túl, a kezdeti keretszintű jellemzőkészleten is hajthatunk végre előfeldolgozást. Előfordulhat, hogy az eredeti adatok túlságosan szétszórva helyezkednek el a térben, valamint vannak köztük olyan minták, melyek kiugró értékeikkel fals irányba mozdíthatják a tanulást. Ennek kiküszöbölésére a jellemzővektorokat előfeldolgozásnak vetettük alá.

Jellemző-transzformáció	a	UAR CV	UAR Teszt	Codebook méret
Normalizálás	5	58,08%	48,13%	512
	10	57,48%	50,27%	512
Standardizálás	5	55,43%	53,54%	512
	10	56,57%	64,32%	256

1. táblázat. Baseline: a keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció során és a teszthalmazon.

A tesztelt értékek az alábbiként alakultak:

- Codebook méretek: 32, 64, 128, 256, 512, 1 024, 2 048
- Kvantáláskor figyelembe vett legközelebbi szomszédok száma: 5, 10
- Előfeldolgozási opciók: standardizálás, normalizálás

4. Tesztek és eredmények

A következőkben ismertetésre kerül a kísérletek pontos menete, valamint az egyes fázisokban kapott eredmények kiértékelése. Mivel minden hangadatbázisból 2×65 darab jellemzőt vontunk ki, hogy figyelembe vehessük a delta értékeket is, így párhuzamosan két külön codebook is készült az alap és a delta jellemzőkhöz az elemzések során. Ebből adódóan, az itt feltüntetett codebook méreteket 2-vel szorozva kapjuk meg az aktuálisan felhasznált jellemzők számát.

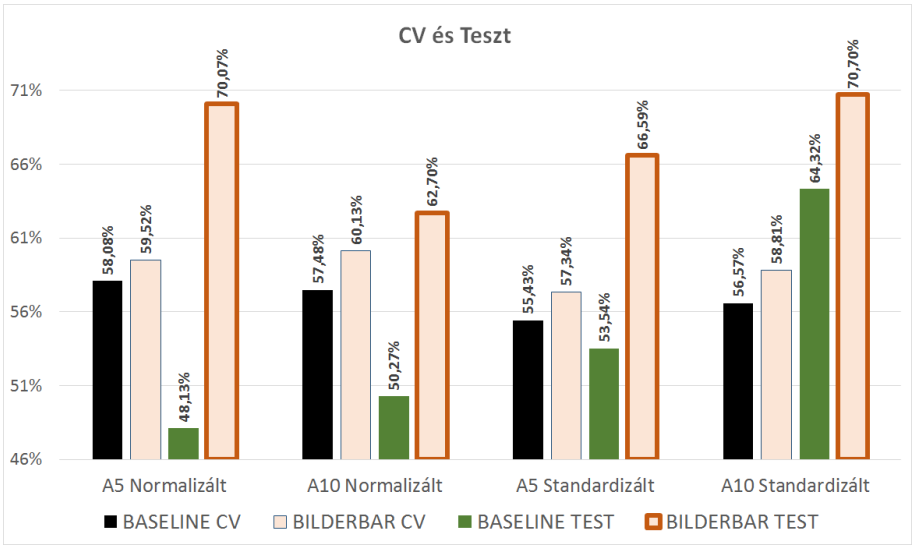
Az egyes tesztesetek közötti különbséget a BoAW reprezentáció előállításához használt különböző codebookok adják. Három esetet vizsgálunk:

- A felhasznált codebook a magyar érzelemadatbázis tanító halmazából készül
- A felhasznált codebook a német érzelemadatbázisból készül
- A felhasznált codebook a magyar általános beszédadatbázisból készül

Első feltevésünk alapját korábbi cikkünk (Vetráb és Gosztolya, 2019) szolgáltatta. Az ott kapott eredmények alapján az akusztikus szózsák jellemzőreprezentációs technika az érzelemdetektálás feladatában versenyképesnek bizonyult. Legnagyobb hátránya a jellemzők előállításához használatos codebook előállításának hosszú ideje volt, így felmerült a kérdés, miszerint egy adott adatbázis jellemzőkészletének kinyeréséhez alkalmasan felhasználható-e másik adatbázisból kapott codebook.

4.1. Baseline

Baseline-ként a magyar nyelvű érzelemorientált felvételeket tartalmazó mintahalmazt használtuk a codebook-ok meghatározásához. A kapott értékek az 1. táblázatban láthatók.



2. ábra: A baseline, valamint az eltérő emocionális adatbázisból származó codebookkal futtatott keresztvalidáció és teszt eredményei

A keresztvalidáció során kapott legjobb eredmény, normalizálással, 5 szomszéd figyelembe vételével, 512 méretű codebookkal és adódott, 58, 08%-ra. A teszt során kapott legjobb eredmény, standardizálással, 10 szomszéd figyelembe vételével, 256 méretű codebookkal adódott, 64, 32%-ra. Ezen kívül megfigyelhető, hogy 4-ből 3 alkalommal a tesztadatbázison kapott eredmények alacsonyabbak voltak, mint a keresztvalidálás eredményei.

4.2. Eltérő, emocionális adatbázisból származó codebook

Első kísérletünkben vizsgáltuk, hogy más adatbázisból előállított codebookkal való munka, képes-e hasonló vagy jobb eredményeket produkálni, mint az eredeti adatbázisból előállított codebook. A codebookokat a korábban már leírt, német nyelvű, érzelem orientált felvételeket tartalmazó, általunk *"bilderbar"*-ként hivatkozott adatbázisból készítettük el. Ezután, ezen codebook-okat felhasználva elkészítettük az akusztikus szózsák jellemzőreprezentációt a magyar nyelvű, érzelem orientált adatbázishoz. A tesztek során kapott értékek az 2. táblázatban láthatók.

A keresztvalidáció során kapott legjobb eredmény, normalizálással, 10 szomszéd figyelembe vételével, 256 méretű codebookkal adódott, 60, 13%-ra. A teszt során kapott legjobb eredmény, standardizálással, 10 szomszéd figyelembe vételével, 256 méretű codebookkal adódott, 70, 70%-ra.

A 2. ábrán látható, hogy mind a négy tesztesetnél szignifikáns javulást értünk el a baseline-hoz képest. Normalizált adat és 10 szomszéd figyelembevételénél,

Jellemző-transzformáció	a	UAR		Codebook méret
		CV	Teszt	
Normalizálás	5	59, 52%	70, 07%	1 024
	10	60, 13%	62, 70%	256
Standardizálás	5	57, 34%	66, 59%	128
	10	58, 81%	70, 70%	256

2. táblázat. BilderbarDB: A keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció és a teszt során.

valamint standardizált adat és 5 szomszéd figyelembevételénél a szükséges codebook méretének csökkenését is tapasztaltuk, melynek hatására a jellemzőreprezentáció előállításához szükséges idő is csökkent. Bár a keresztvalidálás során 4-ből 3 esetben rosszabbul teljesített az eltérő adatbázisú codebookkal dolgozó modell, a tesztek során mégis minden esetben jobb eredményeket kaptunk, mint a baseline. Ez arra enged következtetni, hogy a saját adatbázisból készített codebook túltanulásra hajlamosítja a modellünket, míg az eltérő adatbázisú codebook használata kiegyensúlyozza ezt.

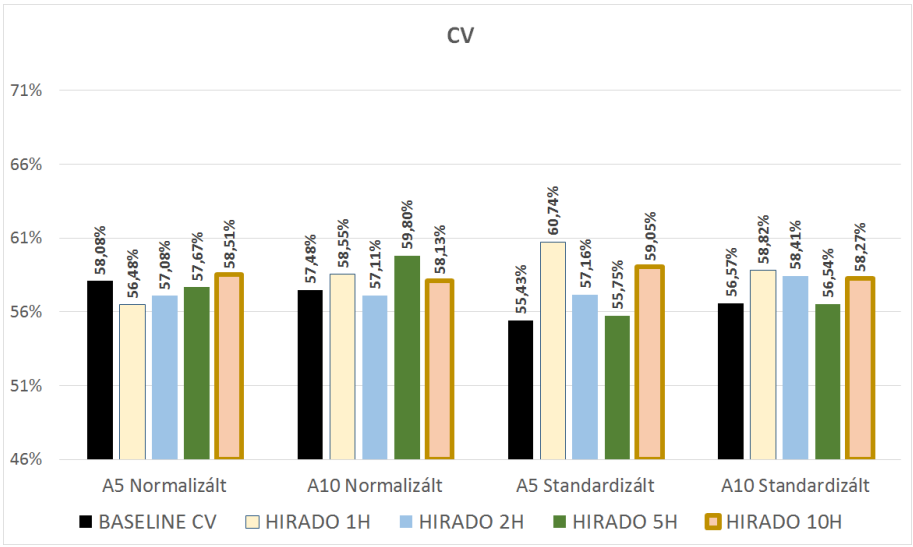
4.3. Eltérő, semleges adatbázisból származó codebook

Az előző tesztek alapján jól látszódott, hogy az eltérő adatbázisból előállított codebook-ok szignifikáns javulást hoztak. Mivel a codebook előállításának fő lépése a felügyelet nélküli klaszterezés, így felmerült a kérdés, hogy vajon befolyásolja-e az osztályozás sikerességét, ha a codebook készítéséhez használt adatbázis eltérő célra készült, mint a codebook-ot felhasználó adatbázis osztályozási feladata?

Az új codebook-okat a korábban már említett, nem érzelemdetektálási célra készült, magyar nyelvű, televíziós híradás felvételeit tartalmazó adatbázis részhalmazaiból állítottuk elő, majd ezek segítségével készítettük el a magyar emocionális adatbázis jellemzőreprezentációját.

Négyféle esetet vizsgáltunk annak érdekében, hogy kiderítsük, vajon a codebook elkészítéséhez használt adatbázis hossza befolyásolja-e az osztályozó teljesítményét: 1 órányi, 2 órányi, 5 órányi és 10 órányi felvételre készítettünk elemzést. A kapott értékek a 4.3. táblázatban láthatók.

A 3. ábrán látható keresztvalidáció során kapott eredményeken semmilyen irányú általános konvergenciát nem tapasztaltunk, így nem tudtunk összefüggést vonni az adatbázis hossza és az osztályozás sikeressége között. Ugyanez mondható el az előfeldolgozási módszer típusáról és a legközelebbi szomszédok számáról is. Minden eredmény 55, 75% és 60, 74% között mozgott. A legjobb eredményeket adó codebook méretek jelentős része 1 024 és 2 048 volt, ami igen nagy jellemző bázist jelent, ami könnyen eredményezhet túlillesztést, így ilyen szempontból a kisebb codebook-okat igénylő német adatbázissal való munka jobbnak bizonyul.



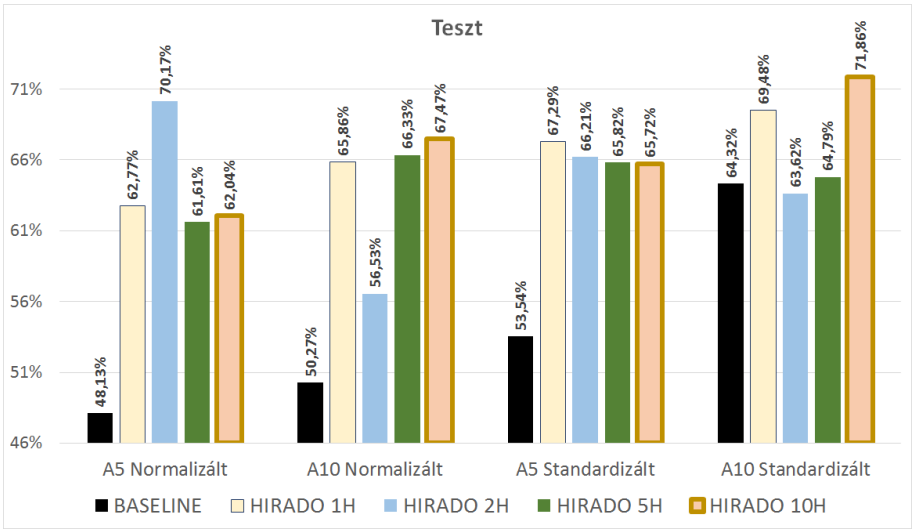
3. ábra: A baseline, valamint az általános magyar adatbázisból származó codebookkal futtatott keresztvalidáció eredményei

Keresztvalidálás során a legjobb eredményt, azaz 60, 74%-ot az 1 órás felvételekkel értük el, standardizálással, 5 szomszéd figyelembe vételével és 1 024-es codebook mérettel. Az eredmények a felvételek hosszától függően, szignifikáns eltérést nem mutattak, így ezzel kapcsolatban nem fedeztünk fel hosszútávú összefüggéseket. Ezen kívül nem kaptunk szignifikánsan se jobb se rosszabb eredményt, mint a kifejezetten érzelemdetektálás céljához készített, német adatbázisból előállított codebook felhasználásakor.

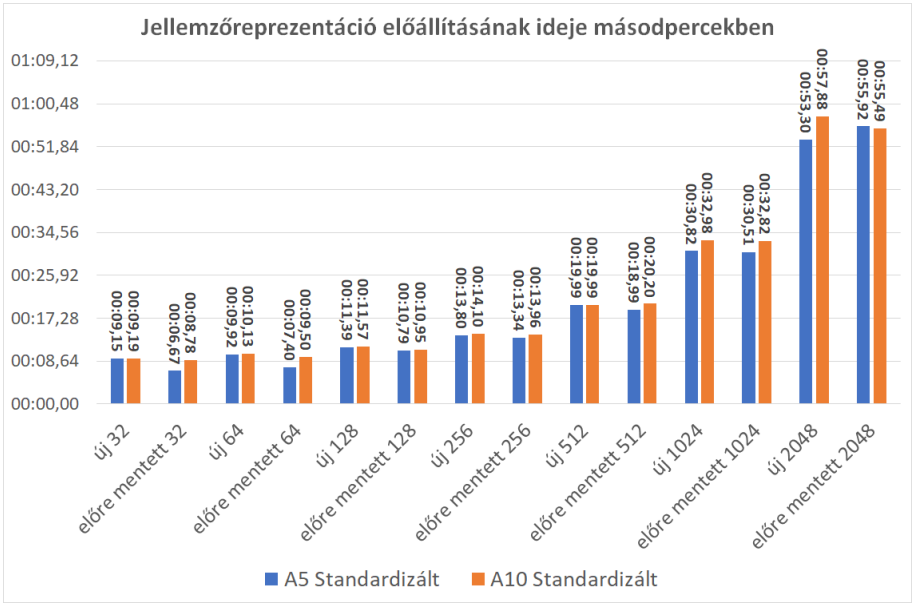
A 4. ábrán feltüntetett, teszthalmazon kapott eredmények nagyon hasonlónak adódtak az előző, német adatbázis segítségével végzett tesztekhez. Itt is elmondható, hogy az eltérő adatbázisból származó codebook használata minden esetben javított a keresztvalidálás eredményein (egy 0, 58%-os visszaesést kivéve a 2 órás adatbázis - 10 legközelebbi szomszéd - normalizálás esetben).

Tesztelés során a legjobb eredményt, azaz 71, 86%-ot, a 10 órás felvételekkel értük el, standardizálással, 10 szomszéd figyelembe vételével és 1 024-es codebook mérettel. Ám a szükséges codebook méretek növekedésén túl további konzekvenciákat itt sem vonhatunk le.

Mivel eredményeink számításához használt adathalmazunk kisméretű, valamint a cikkünkben szereplő statisztikák nem mutatnak teljesen sztochasztikus viselkedést, így felmerülhet a kérdés, miszerint eredményeink mennyire lehetnek véletlenszerűek? Ez azonban egyértelműen kizárható. Egy korábbi tanulmány (Vetráb és Gosztolya, 2019) ugyan ezen adatbázissal végzett tesztjei minden codebook méretre vonatkozóan tartalmazzák az eredményeket, melyek így átfogóbb képet mutatva egyértelműsítik a teljesítmények közötti korrelációkat.



4. ábra: A baseline, valamint az általános magyar adatbázisból származó codebookkal futtatott teszt eredményei



5. ábra: A BoAW jellemzőreprezentáció előállításához szükséges idők a újonnan létrehozott és már meglévő codebook felhasználásának esetében, codebook méret (x tengely) és a használt beállítások függvényében.

Adatbázis	Jellemző-transzformáció	a	UAR CV	UAR Teszt	Codebook méret
1 órás híradó	Normalizálás	5	56, 48%	62, 77%	512
		10	58, 55%	65, 86%	1 024
	Standardizálás	5	60, 74%	67, 29%	1 024
		10	58, 82%	69, 48%	1 024
2 órás híradó	Normalizálás	5	57, 08%	70, 17%	1 024
		10	57, 11%	56, 53%	32
	Standardizálás	5	57, 16%	66, 21%	2 048
		10	58, 41%	63, 62%	2 048
5 órás híradó	Normalizálás	5	57, 67%	61, 61%	2 048
		10	59, 80%	66, 33%	1 024
	Standardizálás	5	55, 75%	65, 82%	128
		10	56, 54%	64, 79%	2 048
10 órás híradó	Normalizálás	5	58, 51%	62, 04%	2 048
		10	58, 13%	67, 47%	1 024
	Standardizálás	5	59, 05%	65, 72%	1 024
		10	58, 27%	71, 86%	1 024

3. táblázat. Híradó: A keretszintű jellemzők normalizálásával és standardizálásával elért legjobb pontosságok a keresztvalidáció során.

Az 5. ábrán láthatók a jellemzőreprezentációk előállításához szükséges idők. Szinte minden esetben több, mint 1 másodperc javulást értünk el, mikor előre elkészített codebook-okat használtunk. Először csekélynek tűnhet ez az érték, de képzeljünk el egy valós idejű rendszert. Egy számítógépes program vagy akár telefonos applikáció esetén, ahol a felhasználói élmény fontos részét képezi a rendszer válaszadási ideje, és ahol sok esetben tized másodpercekben mérik ezt, ott már ez az 1 másodperc is javítja szoftverünk minőségét.

5. Összegzés

Jelen cikkünkben az akusztikus szózsák *(Bag-of-Audio-Words, BoAW)* jellemző-reprezentációs eljárást egy időben, többféle adatbázison alkalmaztuk, érzelemfelismerési feladatra. A különböző esetek kombinációi miatt számos gépi tanulási modellt kellett tanítanunk a különböző paraméter-kombinációkra, így a tesztek futási ideje fontos szempont volt. Mért eredményeink alapján a bemeneti jellemzőket továbbra is érdemes azonos skálára hoznunk normalizálás vagy standardizálás segítségével, ám továbbra sem bizonyult egyértelműen jobbnak egyik a másiknál. Hasonlóan a legközelebbi szomszédok számában sem véltünk felfedezni korrelációt, így továbbra is mind az 5, mind pedig a 10 szomszéd használata hasonlóan jónak bizonyul. A már említett, korábbi (BoAW módszertant vizs-

gáló és a magyar érzelemadatbázissal dolgozó) cikkhez képest az egyes ajánlott codebook méretek idegen adatbázisokból történő codebook kivonás segítségével csökkenthetők lettek. Ebben a tekintetben az idegen, de azonos típusú osztályozási feladatra készült adatbázisból kinyert codebook-ok jobban teljesítettek, mint a más típusú adatbázis codebook-ai.

Tesztjeink alapján egyértelműen kijelenthető, hogy egy-egy előállított codebook eredményesen használható más adatbázisok jellemzőreprezentációjának előállításakor. Azzal kapcsolatban, hogy célszerű-e hasonló célból készített adatbázisok között átvinni a codebookot, vagy bármely két adatbázis között átjárható-e az út, nem találtunk egyértelmű választ. Eredményeink mindkét esetben hasonló skálán mozogtak, szignifikáns különbséget nem adva, csupán a codebook méretekben tértek el, így ez a terület további kutatásokat igényel.

Annak kapcsán, hogy elért eredményeink által merre haladhatunk tovább a későbbiekben, több lehetőség is felmerül. Figyelemre méltó kérdés, hogy a codebook átvitel milyen adatbázis típusok között alakulhat a legeredményesebben. Található-e ebben erősebb összefüggés. Emellett lehetőségünk van más keretszintű jellemzőkészleteket is letesztelni.

Köszönetnyilvánítás

Jelen kutatás eredményei az „Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein” című, EFOP-3.6.3-VEKOP-16-2017-0002 számú projekt támogatásával készültek. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal is támogatta (FK 124413). Gosztolya Gábor kutatásait az MTA Bolyai János ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosító: ÚNKP-19-4-SZTE-51) támogatta.

Hivatkozások

Burkhardt, F., van Ballegooy, M., Engelbrecht, K.P., Polzehl, T., Stegmann, J.: Emotion detection in dialog systems: Applications, strategies and challenges. In: ACII. pp. 985–989. Amszterdam, Hollandia (Sep 2009)

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Interspeech. pp. 1517–1520 (2005)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27 (2011)

Grósz, T., Tóth, L.: A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition. In: TSD. pp. 36–43. Pilsen, Csehország (2013)

Hossain, M.S., Muhammad, G.: Cloud-assisted speech and face recognition framework for health monitoring. Mobile Networks and Applications 20(3), 391–399 (2015)

James, J., Tian, L., Inez Watson, C.: An open source emotional speech corpus for human robot interaction applications. In: Interspeech. pp. 2768–2772. Hyderabad, India (Sep 2018)

Norhafizah, D., Pg, B., Muhammad, H., Lim, T.H., Binti, N.S., Arifin, M.: Detection of real-life emotions in call centers. In: ICIEA. pp. 985–989. Siem Reap, Kambodzsa (June 2017)

Pancoast, S., Akbacak, M.: Bag-of-Audio-Words approach for multimedia event classification. In: Interspeech. pp. 2105–2108. Portland, USA (Sep 2012)

Schmitt, M., Schuller, B.: openXBOW – Introducing the Passau open-source crossmodal Bag-of-Words toolkit. Journal of Machine Learning Research 18(96), 1–5 (2017), http://jmlr.org/papers/v18/17-113.html

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. pp. 148–152 (08 2013)

Sztahó, D., Imre, V., Vicsi, K.: Automatic classification of emotions in spontaneous speech. In: COST 2102. pp. 229–239. Budapest (2011)

Vetráb, M., Gosztolya, G.: Érzelmek felismerése magyar nyelvű hangfelvételekből akusztikus szózsák jellemzőreprezentáció alkalmazásával. In: MSZNY. pp. 265–274. Szeged (2019)

Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Interspeech. pp. 1841–1844. Lisszabon, Portugália (Sep 2005)

A nyelvkontúrkövető algoritmusok és a gépi tanulás összekapcsolhatóságának vizsgálata

Trencsényi Réka

Debreceni Egyetem, Villamosmérnöki Tanszék

Kivonat A publikáció a digitális beszédszintézis tárgykörébe tartozik, és ötvözi a vizuális információkra épülő artikulációs beszédszintézis, illetve a gépi tanulóalgoritmusok eszköztárának alkalmazását. A vizuális információkat dinamikus MRI- és UH-felvételek automatikusan illesztett nyelvkontúrjaiból kinyerve gépi tanítást valósítunk meg, melynek célja a nyelvkontúr hiteles rekonstrukciója. A neurális hálózat be- és kimeneti paramétereinek különböző beállításával módosítható a tanulóalgoritmus jellege. Ennek megfelelően három különböző irányvonal mentén történik tanítás: MRI-adatokból MRI-nyelvkontúrt, UH-adatokból UH-nyelvkontúrt, illetve UH-adatokból MRI-nyelvkontúrt hozunk létre.
Kulcsszavak: artikulációs beszédszintézis, nyelvkontúrkövetés, gépi tanulás

1. Bevezető

A beszédkutatás egyik legdinamikusabban fejlődő, ugyanakkor egyre összetettebb technikai és módszertani kihívásokat rejtő területe a beszédfelismerés mellett a digitális beszédszintézis, ami már napjainkban is szerves részét képezi az ember-gép kapcsolatnak. Ebben a vonatkozásban kulcsfontosságú a gép kommunikációs szerepe, hiszen alapvető rendeltetése a szöveg-beszéd transzformáció megvalósítása, azaz a természetes emberi beszéd közben kialakuló akusztikai produktum élethű utánzása. Ennek kiterjesztett változatában a beszédet jellemző szupraszegmentális elemek (beszédritmus, hangerő, hangmagasság, hangszín, hanglejtés, hangsúly) figyelembevételével tovább finomítható a modell, aminek a beszédfelismerés területén is igen nagy jelentősége lehet (Czap és Pintér, 2015). Napjainkban a beszédszintézis területén zajló kutatások főként a szövegfelolvasó rendszerek megalkotására és tökéletesítésére fókuszálnak, ami olyan alkalmazások elterjedését teszi lehetővé, mint például az utastájékoztató rendszerek, a beszélő okoskészülékek, a szépirodalmi felolvasók, a képernyőolvasók vagy a telefonos tudakozó szolgáltatás. A kutatások hagyományos irányvonalát képviselő szövegfelolvasók esetén a beszédépítés emberi hangminták közvetlen vagy közvetett felhasználásával történik. A törekvések sikerességét a szakirodalom számos közleménye bizonyítja (Olaszy, 1999; Olaszy és mtsai, 2000; Németh és mtsai, 2006; Sproat, 1997; Schröder és Trouvain, 2003; Besacier és mtsai, 2014), melyek igen gazdag tudásanyagot és sokrétű tapasztalatot tükröznek. A klasszikus koncepciók mellett azonban olyan területek is kezdenek kibontakozni, melyek

kevésbé kidolgozottak, és rengeteg nyitott probléma vár még megoldásra. Ide sorolható például az artikulációs (Zappi és mtsai, 2016; Czap és mtsai, 2019) vagy a gépi tanuláson alapuló beszédszintézis (Wu és mtsai, 2015; Arik és mtsai, 2017).

Az artikulációs beszédszintézis az akusztikai produktum utánzását emberi hangminták alkalmazása helyett az emberi hangképzés és artikuláció gépi leképezése révén próbálja megvalósítani. Ennek egyik modern technológiai vonulata a robotok beszédének előállításához szükséges artikulációs elektromechanikus beszédkeltőkre irányuló kísérletezés. A szintézis kiindulópontja az artikulációs-akusztikai konverzió végrehajtása, ami a beszédhez kapcsolódó vizuális információra épül (Czap és Mátyás, 2005). Ennek folytán lényegi szerepet kapnak a különböző képalkotó eljárások (például mágnesesrezonancia-képalkotás (MRI), komputertomográfia (CT), ultrahang (UH)), melyek új információcsatornákat kapcsolnak be a tudományos kutatások folyamatába. Ennek megfelelően a beszéd közben készült MRI- vagy UH-felvételek potenciális forrásai lehetnek az emberi artikulációt jellemző paraméterek vizuális módon támogatott kinyerésének. Mivel a hangok képzésében legaktívabban a nyelv vesz részt, így elsősorban a nyelv mozgását célszerű a lehető legpontosabban monitorozni. Az utóbbi években az erre irányuló vizsgálatok közkedvelt eszközei a már említett MRI, CT és UH mellett az elektropalatográfia (EPG) vagy az elektromágneses artikulográfia (EMA). Az egyszerűbben hozzáférhető UH, EPG és EMA eljárások alkalmazásával csak bizonyos síkmetszetek mentén kaphatunk információt a beszéd dinamikai jellemzőiről, míg a klinikai körülményeket igénylő MRI és CT berendezések segítségével akár háromdimenziós morfológiai adatokra is szert tehetünk. A közelmúltban több tanulmány is foglalkozott dinamikus nyelvkontúr-követési algoritmusok kidolgozásával és fejlesztésével (Li és mtsai, 2005; Csapó és mtsai, 2017; Zhao és Czap, 2019), ami az egyik alappillérét képezheti az artikulációs beszédszintézis témakörében végzett kutatásoknak. A nyelvkontúr dinamikus letapogatását a szagittális síkban érdemes elvégezni, ahol egy kétdimenziós metszeten látható a nyelv fel-le, illetve előre-hátra irányú mozgása. A vizsgálatok legkényelmesebb kellékei UH- vagy MRI-felvételek lehetnek, melyek előnye a jó térbeli és időbeli felbontás, a kép- és hanganyag szinkronizálhatósága, illetve a beszélő alany sugárterheléstől való mentesítése. A nyelvkontúr kijelölése történhet manuálisan vagy automatikus algoritmusok segítségével, bár az adott felvételt alkotó képkockák számának százas vagy akár ezres nagyságrendje indokolttá teszi a dinamikus programozás favorizálását a kézi erővel szemben. A nyelvkontúr detektálásának hatékonyságát nagymértékben meghatározza a felvétel minősége, illetve a kontúrkövető algoritmus típusa (például AutoTrace3, EdgeTrak, TongueTrack, AutoTrace3.5) ezért gyakorlatilag elévülhetetlen ambíció a nyelvkontúrkövető programok finomítása.

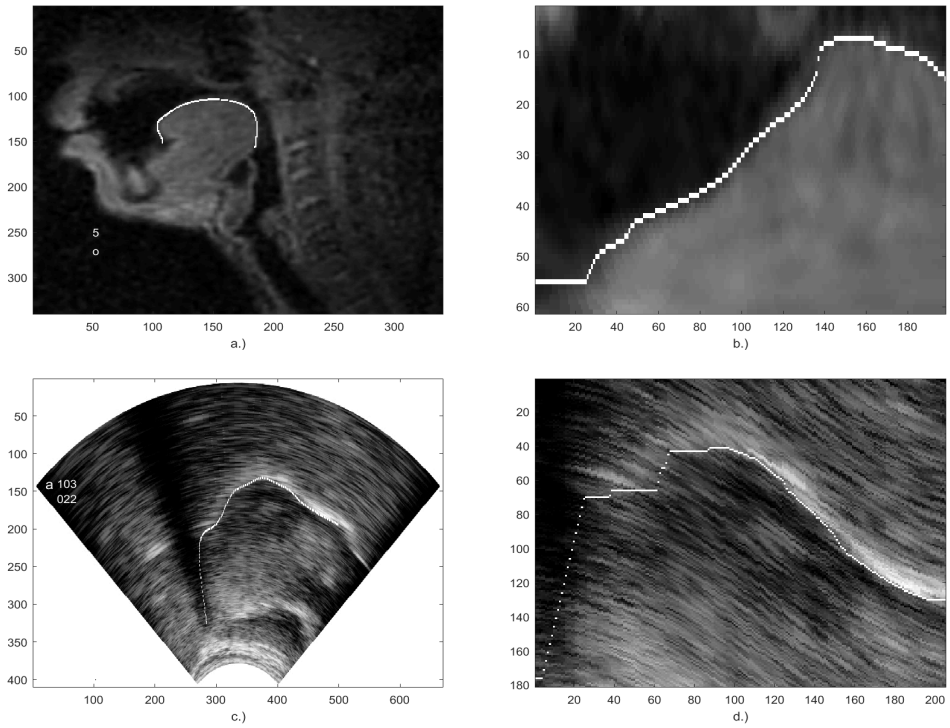
Ezen túlmenően perspektivikus irányvonalat jelöl ki a gépi tanulóalgoritmusok alkalmazása is, melynek során a gép bizonyos bemeneti paraméterek halmazából a környezetéből nyert információk alapján kimeneti eredményeket produkál, miközben javítja a teljesítőképességét. A gépi tanulóalgoritmus lényegében az emberi agy működését próbálja imitálni, így kulcsfontosságú szerepet játszik

a neurális hálózatok működésének ismerete, illetve élethű modellezése. A biológiai neurális hálózatok mintázatok alapján valósítják meg a tanulási folyamatot, ami a gépi tanulás esetében megfelelő algoritmusok megalkotásával képezhető le. A beszédszintézis területén a gép bemeneti paramétereinek halmazát képezhetik például emberi hangminták vagy vizuális forrásokból nyert adatok, melyekkel elvégezve a betanítást megszólaltatható az auditív produktum. A vizuális információkkal betanított neurális hálózat lehetősége tehát természetes módon kínálja fel az artikulációs beszédszintézis és a gépi tanulás módszereinek összekapcsolását. A lehetőségek jóformán korlátlanok, az eljárások, illetve ezek kombinációja pedig javarészt még nincs kimerítően feltárva. Jelen publikáció a nyelvkontúrkövetés és a gépi tanulóalgoritmusok együttes alkalmazhatóságának bizonyos vonatkozásait vizsgálja MRI- és UH-felvételek feldolgozásával.

2. Automatikus nyelvkontúrkövetés

A vizsgálatok tárgyát beszéd közben készült MRI- és UH-felvételek képezték. Az MRI-felvételeket a Dél-kaliforniai Egyetem honlapján szabadon hozzáférhető multimédiás csomagból válogattam ki, az UH-felvételek pedig az MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoport SonoSpeech rendszerével készült audiovizuális anyagok formájában álltak rendelkezésemre.

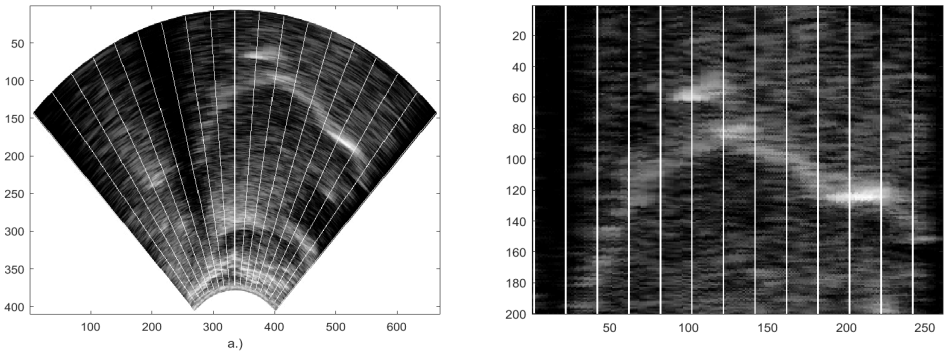
A nyelvkontúrkövetés elsődleges célja a beszédhangokhoz tartozó nyelvállások dinamikus leírása, illetve a koartikuláció során létrejövő hangátmeneteket jellemző nyelvmozgások tanulmányozása. A kvalitatív analízis mellett a nyelvkontúr a beszéd kvantitatív jellegű tanulmányozásának is jó kiindulópontja lehet, hiszen a nyelvkontúrból származtatható számszerű értékek elősegíthetik az artikulációs modellek mélyebb megértését és fejlesztését. A nyelvkontúr detektálására kidolgozott algoritmusok az alkalmazott eljárásoktól függően rendkívül változatosak lehetnek. A vizsgálataim segédeszközeként olyan algoritmust használtam fel, amely a dinamikus programozás technikáját alkalmazza. A nyelvhát az UH-felvételen világos sávként rajzolódik ki, az MRI-felvételen pedig a szájüregi levegő sötét tartománya és a nyelvszövet világos tartománya között létrejövő kontrasztként érzékelhető, így a kontúrkövetés mindkét esetben a nyelvhát vonalát meghatározó maximális világosságú képpontok megkeresését jelenti. Az algoritmus alkalmazását a felvételek előfeldolgozása előzi meg, ami a képalkotó eljárásokból adódó zajok és folytonossági hiányok megszüntetésére irányul. Az említett hibák redukálásának leghatékonyabb eszközei az élkiemelő és átlagoló operációk, amik matematikailag a konvolúció műveletével valósíthatók meg (Czap, 2007). A megkeresett maximális világosságú képpontok, igazodva a nyelvhát egyenetlen vonalához, egy nyers görbét hoznak létre, melynek simítása diszkrét koszinusz transzformációval oldható meg. Az 1. ábra képei automatikusan illesztett nyelvkontúrt mutatnak be egy-egy MRI- (a.) és b.)), illetve UH-kereten (c.) és d.)). Az 1.a ábrán az *o* hanghoz tartozó nyelvállás figyelhető meg, míg az 1.c ábra az *a* hangnak megfelelő nyelvállást jeleníti meg a simított nyelvkontúr kiemelésével. Az 1.b és 1.d ábrákon az 1.a, illetve 1.c kereteken megrajzolt nyelvontúrok simítás nélküli, kinagyított részletei láthatók.



1. ábra: A nyelvkontúr követése MRI- és UH-felvételeken

Az 1.b és 1.d ábrák speciális transzformációval hozhatók létre az 1.a és 1.c ábrákból kiindulva. A transzformációs eljárás lényegét a 2. ábrán látható UH-keret segítségével érzékeltetem. Első lépésként a radiális geometriájú 2.a képen a kör középpontjából kiindulva sugárirányú metszeteket képzünk a felvétel által definiált -45° – 45°-os tartományban, melyek mentén lényegében újramintavételezzük a képet. Az így létrejövő metszeteket oszlopdiagramba rendezzük, melynek eredményeképpen egy olyan képmátrixot kapunk, ami a descartes-i x-y síkban jellemezhető a legkényelmesebben. A mátrixos szerkezet kialakítása nyomán áll elő a 2.b ábra. A vizsgálatok azt mutatják, hogy az 1/4°-onként végrehajtott mintavételezés a legideálisabb, hiszen ekkor a mátrix szomszédos oszlopai között nem fordul elő két pixelnél nagyobb változás a kontúrban. Az áttekinthetőség kedvéért a metszeteket csak 5°-onként ábrázoltam, amit a 2. ábra fehér vonalai szemléltetnek. Az eljárás MRI-keretek esetében ugyanilyen módon működik az MRI-kereten megfelelően kijelölt középpont és (a -45° – 45°-os tartománytól általában szélesebb) szögtartomány alkalmazásával.

Az MRI-felvételek adatközlője angol anyanyelvű férfi beszélő, aki VCV típusú hangsorokat szólaltat meg V magánhangzóval és C mássalhangzóval. A bemu-



2. ábra: Radiális és mátrixos geometriájú UH-keretek

tatott MRI-keret tanúsága szerint a kapott görbe hitelesen követi a nyelvhát vonalát. Az UH-felvételeken magyar, illetve kínai anyanyelvű női bemondótól származó hangsorok vannak rögzítve, melyek CVC, illetve VCV szerkezetűek. Összehasonlítva az 1. ábra képeit, feltűnhet, hogy az UH-felvételen a nyelvhát kevésbé éles határvonalként jelenik meg, ami egy elmosódott világos sávot eredményez. Ez a nyelv és a fölötte lévő levegő határán visszaverődő UH-hullámok következményeként alakul ki, így a nyelvkontúr a világos sáv alsó határán lokalizálható. Az UH-felvételek további sajátsága, hogy a nyelvgyök és az állcsont árnyékoló hatása miatt a nyelv hátsó része és a nyelvhegy nem látható a felvételen, így a nyelv alakjáról és mozgásáról csak részleges információt kaphatunk. A nyelvgyök és az állcsont árnyéka sötét sávként azonosítható az 1.c ábra bal és jobb oldali részén.

3. Gépi tanulás

Jelenlegi kutatómunkám célkitűzése az előző fejezetben bemutatott nyelvkontúrkövetés és a gépi tanulóalgoritmusok összekapcsolása, illetve az egymáshoz való viszonyuk bizonyos aspektusainak tanulmányozása. Programjaimat MATLAB-környezetben hoztam létre, és a gépi tanítást olyan algoritmussal valósítottam meg, amely a neurális hálózat súlyfaktorait a skálázott konjugált gradiens módszer (Moller, 1993) segítségével határozza meg. Ezen optimalizációs eljárás a problémához rendelt egyenletrendszert a bemeneti paraméterek ismeretében iterációs módszerrel oldja meg, miközben az eljárással számított kimeneti paraméterek értékei konvergálnak az előírt értékekhez. A módszer előnye, hogy az iterációs algoritmus lépésközeinek számát minimalizálva elég gyors konvergencia biztosítható, így a gépi tanítás viszonylag rövid idő alatt véghezvihető. Az iterációs lépések olyan irány mentén valósulnak meg, ami gyorsabb konvergenciát biztosít, mint a legmeredekebb ereszkedésnek megfelelő legnegatívabb gradiens, miközben megőrzi a korábbi lépésekben kapott hibaminimalizációt.

A neurális hálózatban két rejtett réteget helyeztem el, melyek egyenként 30 neuront tartalmaztak. A tanuláshoz szükséges bemeneti paramétereket a dinamikusan változó nyelvkontúr négy kiválasztott pontjának segítségével jelöltem ki, melyekhez kimeneti paraméterként a nyelvkontúr diszkrét koszinusz transzformáltját rendeltem hozzá. A négy kiválasztott pont relatív helyzete minden képkockán azonos olyan értelemben, hogy a négy pont minden nyelvkontúr esetében a görbe hosszának kb. 20%, 40%, 60%, 80%-ánál található.

A tanítást elsőként az MRI-forrásból származó be- és kimeneti paraméterek rögzítésével hajtottam végre, az eredményeket pedig ugyanazon MRI-kereteken teszteltem. A procedúrát hasonló elv alapján az UH-keretekre is megismételtem, végül az UH-forrásból kinyert bemeneti paraméterek, illetve az MRI-forrásból eredő kimeneti paraméterek kombinálásával újra lefuttattam az algoritmust, majd eredményeimet az MRI-kereteken teszteltem. A következő alfejezetek a három különböző megközelítést tárgyalják.

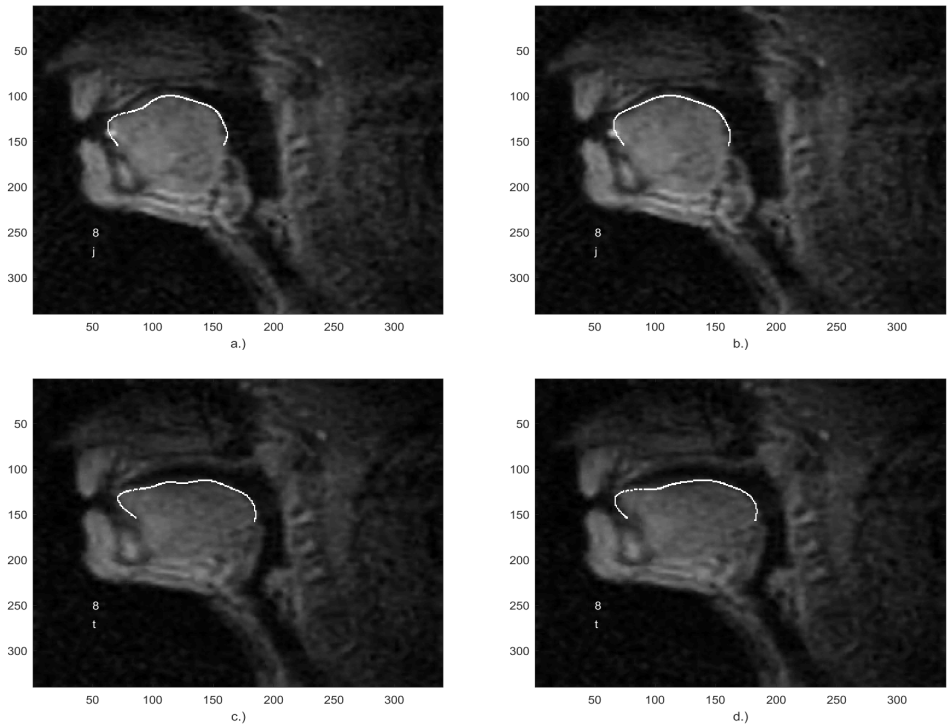
3.1. MRI-nyelvkontúr tanítása MRI-adatokkal

Az alfejezet az MRI-felvételek esetében elvégzett gépi tanítás eredményeit foglalja össze. A tanítás alapját az *a,á,c,cs,d,dz,dzs,e,é,g,i,j,k,l,n,o,ö,r,s,sz,t,u,ü,z,zs* beszédhangokhoz tartozó fonemikus konfigurációk képezték. A bemeneti paramétereket a nyelvkontúr négy kiválasztott pontjának képsíkban mért y koordinátája adta, a kimeneti paraméterek halmazát pedig a nyelvkontúr diszkrét koszinusz transzformáltjának első húsz együtthatója határozta meg, melynek alapján a tanulóalgoritmus futtatását követően inverz diszkrét koszinusz transzformációval rekonstruálható a betanított nyelvkontúr. Ez lényegében azt jelenti, hogy mindössze négy pont felhasználásával történik a teljes görbe előállítása. Eredményeimet a *j* és *t* hangok példáján keresztül mutatom be.

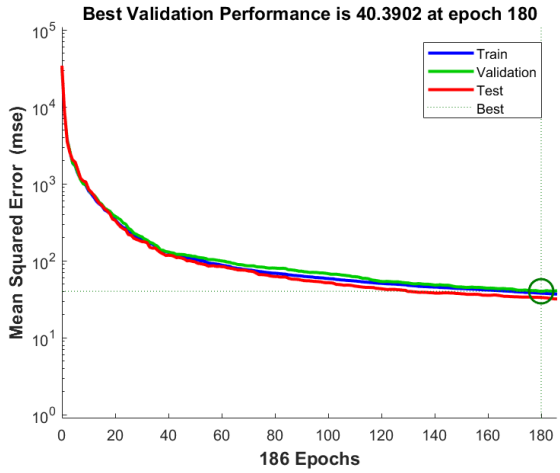
A 3.a és 3.c ábrák a *j*, illetve a *t* hangnak megfelelő nyelválláshoz illesztett nyelvkontúrokat prezentálnak. A 3.b és 3.d ábrák ugyanazon *j*, illetve *t* hanghoz tartozó betanított nyelvkontúrokat jelenítenek meg. Az illesztett és a betanított nyelvkontúrok összehasonlításakor nem mutatkozik figyelemreméltó vizuális különbség, minimális az eltérés a két görbe között. A 3. ábrán szemléltetett eredmények azt tükrözik, hogy a tanulóalgoritmus hatékonyan működik, amit a 4. ábra grafikonjai is alátámasztanak. Az ábrán a tanítás, a tesztelés és a validálás átlagos négyzetes hibája követhető nyomon. Látható, hogy gyors csökkenés mellett a tanítás és a tesztelés hibája lényegében azonos.

3.2. UH-nyelvkontúr tanítása UH-adatokkal

Az alfejezet az UH-felvételek esetében elvégzett gépi tanítás eredményeit foglalja össze. A tanítás ez esetben a "Most a CVCV, meg a CVCV volt." típusú bemondásokra épült. A bemeneti és kimeneti paraméterek értelmezése ugyanaz, mint az előző alfejezetben, és a lépéseket ezúttal a *g* és *s* hangok példáján keresztül vezetem végig.

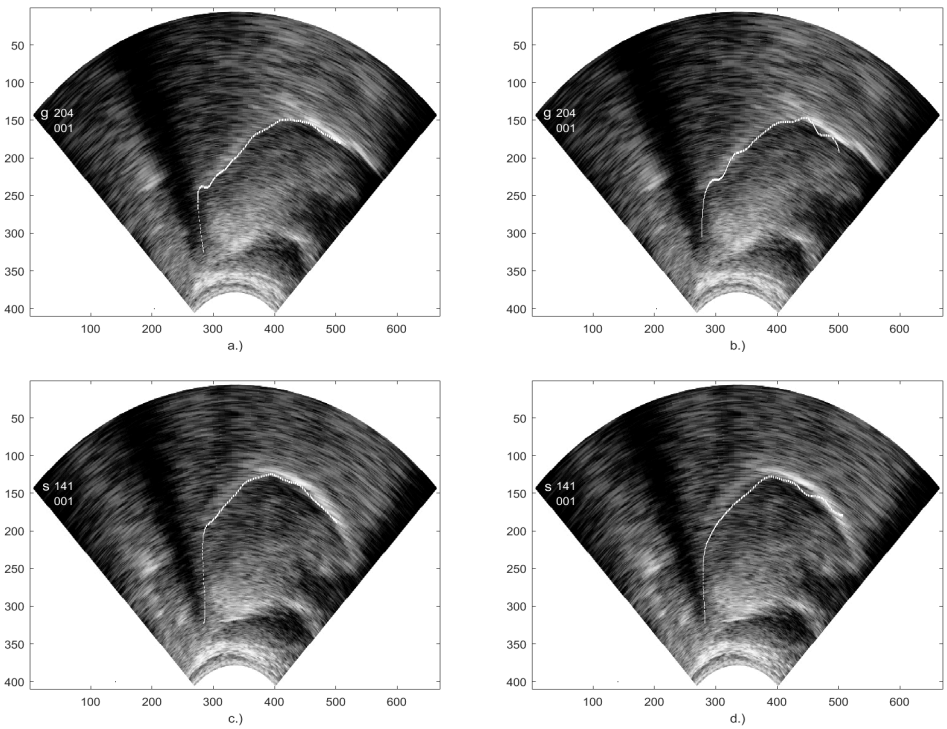


3. ábra: Az illesztett és betanított MRI-nyelvkontúr a *j* és *t* hangok esetében



4. ábra: A gépi tanítás átlagos négyzetes hibája MRI-MRI tanítás esetén

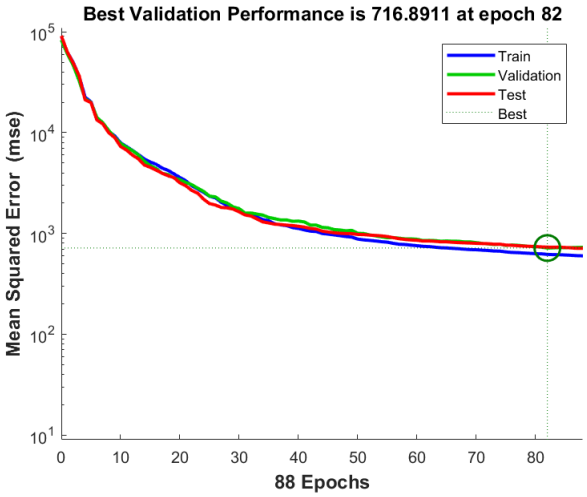
A 5.a és 5.c ábrák a *g*, illetve a *s* hangnak megfelelő nyelválláshoz illesztett nyelvkontúrokat demonstrálnak. A 5.b és 5.d ábrák ugyanazon *g*, illetve *s* hanghoz tartozó betanított nyelvkontúrokat mutatnak be. Összehasonlítva az illesztett és a betanított nyelvkontúrokat, ez esetben sem figyelhető meg számottevő különbség a két görbe között. A tanítás, a tesztelés és a validálás átlagos négyzetes hibájának alakulását az 6. ábra tünteti fel, melynek tendenciája hasonló az MRI-felvételekkel megvalósított tanítás során kapott görbékhez.



5. ábra: Az illesztett és betanított UH-nyelvkontúr a *g* és *s* hangok esetében

3.3. MRI-nyelvkontúr tanítása UH-adatokkal

Az előző két alfejezetben a gépi tanulás be- és kimeneti paraméterei ugyanazon forrásból származtak, hiszen MRI-nyelvkontúrt MRI-adatokkal, UH-nyelvkontúrt pedig UH-adatokkal tanítottunk. Érdemes azonban azt is tanulmányozni, hogy milyen sikerrel kapcsolhatók össze a két különböző forrás paraméterei. Ebből a célból a neurális hálózatot úgy szerkesztettem meg, hogy bemeneti paramétereit az UH-nyelvkontúr négy kiválasztott pontja, kimeneti paramétereit pe-

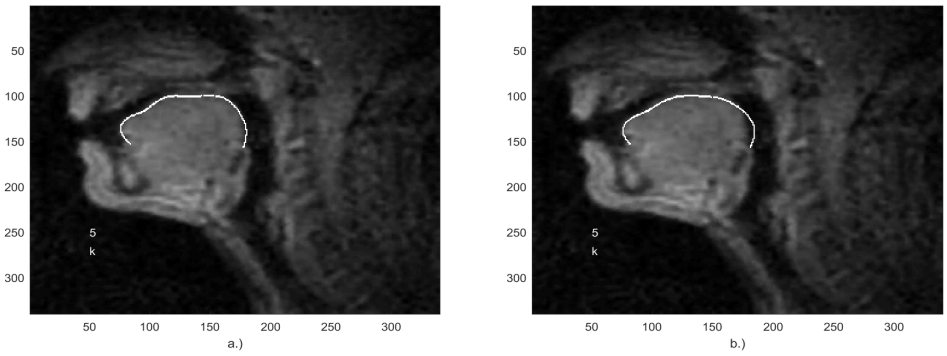


6. ábra: A gépi tanítás átlagos négyzetes hibája UH-UH tanítás esetén

dig az MRI-nyelvkontúr diszkrét koszinusz transzformáltja alkotta. Ezáltal egy olyan tanítási mechanizmus hozható létre, melynek során MRI-nyelvkontúrt alkothatunk UH-adatok felhasználásával. Eredményeim ismertetéséhez újfent az *a* hangot hozom fel példaként. Megjegyzem, hogy a felhasznált adatbázis mérete nagyságrendekkel elmarad a 3.1., illetve 3.2. alfejezetekben taglalt körülményekhez képest. Ennek oka, hogy az MRI- és UH-forrásokból származó felvételek nem minden esetben azonos típusú bemondásokat szólaltatnak meg, és emellett az egyes beszédhangokhoz rendelt képkockák száma sem egyezik meg, ami megnehezíti a tanulóalgoritmus paramétereinek összehangolását. A bemondások és mintaszámok szinkronizálása azonban jelenleg is folyamatban van.

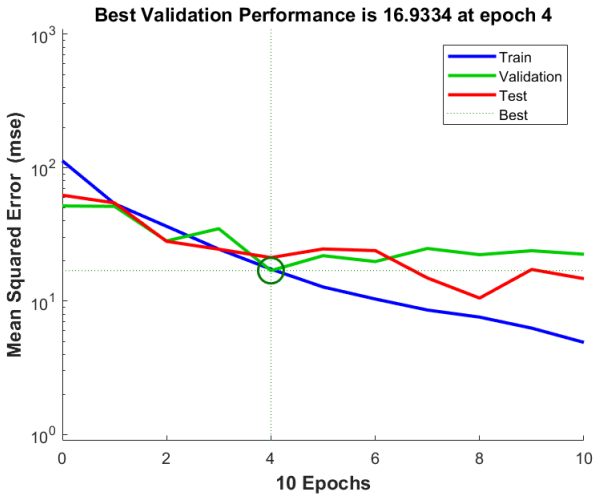
A 7.a ábra a *k* hangnak megfelelő nyelválláshoz illesztett nyelvkontúrt szemléltet. A 7.b ábra ugyanazon *k* hanghoz tartozó betanított nyelvkontúrt illusztrál. Az eredmény akár több szempontból is érdekes lehet, hiszen amellett, hogy különböző anyanyelvű, eltérő nemű adatközlők különböző képalkotó eljárással készített felvételeiből származnak a neurális hálózat révén összekapcsolt be- és kimeneti paraméterek, az sem elhanyagolható körülmény, hogy a tanítás szűkebb adathalmazból kiindulva produkál bővebb adathalmazt. A 2. fejezet végén ugyanis említettem, hogy az UH-felvétel nem képes megjeleníteni a nyelv hátsó részét és a nyelvhegy régióját, ami az MRI-felvételen természetesen akadályok nélkül látható. Ez pedig azt vetíti előre, hogy az UH-felvételekből származó részleges adatokkal tanulóalgoritmusok bevetésével hatékonyan becsülhető a teljes nyelvhát kontúrja.

A 8. ábrán a tanítás, a tesztelés és a validálás átlagos négyzetes hibájának futása elevenedik meg. Látható, hogy a tanítás és a tesztelés hibagörbéje nem mutat olyan mértékű együtthaladást, mint amit a 4. és 6. ábrák tükröznek. Ez a tanítóalakzatok fentebb említett csekély számának a következménye, a kezdeti



7. ábra: Az illesztett és betanított UH-nyelvkontúr a *k* hang esetében

adathalmaz bővítésével azonban javulás várható a görbék relatív lefutásának tekintetében.



8. ábra: A gépi tanítás átlagos négyzetes hibája UH-MRI tanítás esetén

4. Összefoglaló

A cikk az artikulációs beszédszintézisben fontos szerepet játszó automatikus nyelvkontúrkövető algoritmusok, illetve a gépi tanítás együttes alkalmazását demonstrálja dinamikus MRI- és UH-felvételek feldolgozásával. A gépi tanulás a

neurális hálózat be- és kimeneti paramétereinek megfelelő kombinálásával MRI-MRI, UH-UH, illetve UH-MRI viszonylatban valósul meg. Megjegyzem, hogy a jelenlegi fázisban még igen korlátozott számú tanító- és tesztelőalakzat áll rendelkezésre, de a forrásadatok fokozatos bővítés alatt állnak. Az aktuális eredmények a folyamatban lévő kutatómunkából csupán egy keskeny szeletet, egy pillanatképet villantanak fel, hiszen az artikulációs beszédszintézis és a gépi tanulás területei önmagukban véve is rendkívül sok problémát vetnek még fel, amiknek jó része egyelőre csak részlegesen tekinthető megoldottnak. Ennek megfelelően a kutatások jövőbeli irányát meghatározhatja például a vizuális információkra alapozott, statisztikai elven működő vagy szabályalapú algoritmusokkal létrehozott beszédszintézis modelljeinek tökéletesítése, ami alapvető fontosságú lehet például a klinikai célú beszédterápiában, a nem anyanyelvi nyelvtanulási tréningek kialakításában vagy a néma beszéd megszólaltatásához szükséges szintetizátorok konstrukciójában és fejlesztésében.

Hivatkozások

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Andrew, N., Raiman, J., Sengupta, S., Mohammad, S.: Deep voice: Real-time neural text-to-speech. In: Proceedings of the 34th International Conference on Machine Learning, 70, 195-204 (2017)

Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Comm., 56, 85-100 (2014)

Czap, L., Mátyás, J.: Virtual announcer. Infocommunications Journal, 60, 2-5 (2005)

Czap, L., Mátyás, J.: Virtual speaker. In: Ádám, T., Vásárhelyi, J., Varga, A. (szerk.): Proceedings of 6th International Carpathian Control Conference ICCC 2005 Miskolc, Magyarország: Miskolci Egyetem, 351-358 (2005)

Czap, L.: Képfeldolgozás. Miskolc-Egyetemváros, Magyarország: Miskolci Egyetem, 151 p. (2007)

Czap, L., Pintér, J. M.: Intensity feature for speech stress detection. In: Petras, I., Podlubny, I., Kacur, J., Vásárhelyi, J. (szerk.): Proceedings of the 16th International Carpathian Control Conference Miskolc, Magyarország: IEEE IAS/IES/PELS, 91-94. (2015)

Czap, L., Pintér, J. M., Baksa-Varga, E.: Features and Results of a Speech Improvement Experiment on Hard of Hearing Children. Speech Communication, 106, 7-20 (2019)

Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A.,, Varjasi, G.: Szinkronizált beszéd- és nyelvultrahang-felvételek a Sono-Speech rendszerrel. In: Vincze V. (szerk.): XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017). Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 339-346 (2017)

Li, M., Kambhamettu, C., Stone, M.: Automatic contour tracking in ultrasound images. Clinical linguistics and phonetics, 19, 545-554 (2005)

Moller, M. F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural networks, 6, 525-533 (1993)

Németh, G., Olaszy, G., Fék, M.: Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. Beszédkutatás, 183-196 (2006)

Olaszy, G.: Beszédadatbázisok készítése gépi beszédelőállításhoz. Beszédkutatás99, 68-89 (1999)

Olaszy, G., Németh, G., Olaszi, P., Kiss, G.: Profivox: a legkorszerűbb hazai beszédszintetizátor. Beszédkutatás 2000, 167-179 (2000)

Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. Int. J. Speech Tech., 6, 365-377 (2003)

Sproat, R. W.: Multilingual text-to-speech synthesis. KLUWER Academic Publishers (1997)

Zappi, V., Vasuvedan, A., Allen, A., Raghuvanshi, N., Fels, S.: Towards real-time two-dimensional wave propagation for articulatory speech synthesis. In: Proceedings of Meetings on Acoustics 171ASA, 26, 045005 (2016)

Zhao, L., Czap, L.: A nyelvkontúr automatikus követése ultrahangos felvételeken. Beszédkutatás, 27, 331-343 (2019)

Wu, Z., Valentini-Botinhao, C., Watts, O., King, S.: Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 4460-4464 (2015)

ASR-hibaterjedés vizsgálata a gépi beszédértés szemszögéből

Tündik Máté Ákos, Szaszák György

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{tundik,szaszak}@tmit.bme.hu

Kivonat Napjainkban a mesterséges intelligencia alapú megoldások egyre inkább a beszélt nyelv gépi megértésére törekednek. Ennek preferált megközelítése az, amikor automatikus beszédfelismerő (ASR) rendszerek használatával átiratokat hozunk létre, amelyek további, szövegalapú elemzésen mennek keresztül. A gépi átiratok szóhibákat is tartalmazhatnak; ezen hibák továbbterjednek a szöveges feldolgozási folyamatba, így a gépi központozásba, kivonatolásba is. Ugyanakkor szubjektív tesztjeink azt igazolták, hogy az emberek a gépi átiratokat a szóhibák és a központozási hibák ellenére is jól tudják értelmezni. Célunk az, hogy bemutassuk az ASR-hibaterjedésből adódó, szemantikai térben bekövetkező információveszteségeket, valamint az ASR-hibaterjedés automatikus összefoglalásra gyakorolt hatását is elemezzük. Bemutatjuk, hogy az egyes mondatreprezentációk a szóhibák hatására enyhén eltolódnak a szemantikai térben, de ez jócskán elmarad a dokumentum mondatainak átlagos szemantikai távolságától. Megmutatjuk azt is, hogy a központozás hibáinak nagyobb hatása van az összefoglalók kiértékelésére, mint a szóhibáknak, ami arra enged következtetni, hogy a feladathoz elengedhetetlen a megfelelő mondatszintű tokenizálás.
Kulcsszavak: szemantikai hasonlóság, hibaterjedés, gépi beszédértés, tartalmi összefoglalás

1. Bevezetés

A legmodernebb, beszédalapú összefoglaló rendszerek egy automatikus beszédfelismerő (ASR) eszköz segítségével szöveges átiratot készítenek, majd következő lépésként a szöveges dokumentum összefoglalása következik. Az utóbbi modul általában először mondatszintű tokenizálást hajt végre, majd ezt további, a szemantikai térben elvégzett műveletek követik. Az összefoglaló készítése kétféleképpen történhet; (1) ún. extraktív módon, amikor a gépi átirat mondatai kerülnek felhasználásra rangsorolást követően (Celikyilmaz és Hakkani-Tür, 2011), (2) absztraktív módon, amikor egy szemantikai kódolási algoritmus biztosítja a még tömörebb, 'újrafogalmazott' összefoglaló létrehozását (Genest és Lapalme, 2012; Paulus és mtsai, 2017). A szemantikai térbe történő projekció leggyakoribb módja a szóbeágyazások (szóvektorok) használata (Mikolov és mtsai, 2013b).

Az ASR kimeneten átadott átiratok feldolgozásakor központozási hibákkal és szóhibákkal is számolni kell; ezek a hibák továbbterjednek a feldolgozási folyamatban, így befolyásolják a beszéd tartalmi összefoglalását is.

Az első szövegfeldolgozási lépésként elvégzendő mondatokra bontás nehézsége, hogy az írásjelek és a nagybetűk hiányoznak a nyers ASR-átiratokból. A központozás megvalósítására vagy prozódiai alapú szegmentálást végzünk el, közvetlenül a beszédanyagon (Beke és Szaszák, 2016), vagy a gépi beszédátiratban állítjuk vissza az írásjeleket (Klejch és mtsai, 2017; Öktem és mtsai, 2017; Tündik és Szaszák, 2018). Az utóbbi megközelítés alkalmazásával nemcsak akusztikus, hanem nyelvi (szöveges) jellemzők is kiaknázhatók. A legkorszerűbb automatikus központozó rendszerek teljesítménye F1-mértéket tekintve 70-80%, tehát ezen megoldások esetében is még jócskán jelen vannak központozási (írásjelezési) hibák a központozott átiratban.

ASR vonatkozásában - feladattól és a környezeti feltételektől függően - az ipari hasznosítás szempontjából releváns alkalmazásokban a szóhibaarány (WER) 1-30% között van. Kevés tanítóanyaggal rendelkező, vagy nyelvi szempontból tekintve speciális nyelv - például morfológiailag vagy összetett szavakban gazdag, stb. - esetében a WER sokkal magasabb lehet, mint hasonló funkcionalitást nyújtó angol nyelvű alkalmazások esetén. A felhasználói élmény ugyanakkor általában kevésbé romlik le, mint azt a WER különbsége sugallná, sőt, ugyanazon mértékű szóhibaaránnyal működő angol ASR rendszert akár a végfelhasználók rosszabbra is értékelhetnek, mint egy finn (Kurimo és mtsai, 2006) vagy magyar (Tündik és mtsai, 2018) rendszert.

Valójában az emberek meglepően jól teljesítenek, ha hibákkal terhelt, automatikusan központozott gépi átiratokat kell olvasniuk és értelmezniük (Tündik és mtsai, 2018). Nyilvánvaló, hogy a gépi értelmezéssel szemben az emberek tágabb kontextusra és egyéb olyan aspektusokra is támaszkodhatnak, amelyek a gyakran nem is tudatosuló hibajavító mechanizmus működését segítik (Postma, 2000; Kröger és mtsai, 2016). A halláskárosodásban szenvedő személyek esetében korábban igazoltuk, hogy az ép hallású emberekhez viszonyítva jobban teljesítenek a szó-, és különösen az írásjelek hibáinak spontán javításában (Tündik és mtsai, 2018), valószínűsíthetően az ilyen hibák tudatos észlelésének küszöbértéke sokkal magasabb az esetükben.

A szemantikus térbe történő transzformációk, különösen a szóbeágyazások (Mikolov és mtsai, 2013a) nagyon népszerűvé váltak a természetes nyelvi feldolgozásban és a beszélt nyelv megértésében. Noha az ilyen szóvektor-ábrázolások a szemantikai vagy a szintaktikai konzisztencia és pontosság szempontjából messze nem tökéletesek, kiváló képességeket mutatnak az információ szemantikai feldolgozását magában foglaló (pl. következtetési, analógiai) feladatok esetében. A szóvektorok használata korszerűnek számít a tartalmi kivonatolásban is. Jelen cikkünkben az inspirált minket, hogy objektív mérések alapján felmérjük, mennyire torzul az információ a szemantikai térben a szó- és/vagy központozási hibák miatt az automatikus beszéd-szöveg átalakítást következtében. A szemantikai torzítást eddig elsősorban szubjektív szempontból vizsgálták (Kafle és Huenerfauth, 2016; Tündik és mtsai, 2018), ekkor az ASR-hibaterjedésének hatása

a szemantikai térben csekélynek mondható, ésszerű, ipari alkalmazást lehetővé tévő szóhibaarány mellett. Egyes kutatók megvizsgálták a szóhelyettesítési hibák hatását mondatbeágyazások szintjén (Voleti és mtsai, 2018), más munkák (pl. (Simonnet és mtsai, 2018)) az ASR hibák szimulációját javasolták az ilyen típusú elemzésekhez. Mivel a valós ASR átiratok előállítása nem bonyolult, amennyiben a hanganyag rendelkezésre áll, ezért nem szimuláltunk ASR hibákat, hanem valódi gépi átiratokat használtunk, ezzel is kiküszöbölve a szimulációval bevitt torzítást. Ezáltal lehetőségünk nyílt a helyettesítési hibák kizárólagos vizsgálata helyett az összes lehetséges szóhibát számításba venni (így a törléseket és a beszúrásokat is), csakúgy, mint a központozási hibákat, hogy a kísérleti beállítások a lehető legközelebb kerüljenek a valódi felhasználási helyzethez, körülményekhez.

Cikkünk a következőképpen épül fel: bevezetőnkben bemutattuk az ASR-hibaterjedés problémakörének jelentőségét, kifejtettük motivációnkat, és bemutattunk néhány, a témához kapcsolódó munkát. A következő fejezetek a felhasznált adatbázist, valamint a mondatszintű és a dokumentumszintű szemantikai hasonlóság méréséhez használt módszertant dokumentálják, az utóbbihoz egy népszerű, dokumentum-összefoglaló alapú megközelítést használva. Ezt követően bemutatjuk és megvitatjuk eredményeinket, mielőtt végső következtetéseinket levonnánk.

2. Adat, ASR és Központozás

2.1. Átiratok előkészítése

Kutatásunk során az ASR- és/vagy írásjelhibák által okozott szemantikai torzításokat vizsgáljuk. Ezáltal négy különböző, ámbár összehasonlítható átiratváltozatot készítettünk minden egyes beszédfájlra, az alábbiak szerint[1]:

MT-MP: Kézi Átirat - Kézi Központozás : emberek által készített referenciaátirat, amely az alábbi négy írásjelet tartalmazza: {. , ? !};

AT-MP: Gépi (ASR) Átirat - Kézi Központozás : gépi átirat felhasználása, melybe a referenciaátirat segítségével „visszacsempésztük" az írásjeleket[2];

MT-AP: Kézi Átirat - Automatikus Központozás: a referenciaátiratból eltávolítottuk az írásjeleket, majd azokat automatikus módszerrel prediktáltuk (Tündik és mtsai, 2018);

AT-AP: Gépi (ASR) Átirat - Automatikus Központozás: a gépi átiratok automatikus központozásához szintén a (Tündik és mtsai, 2018) cikkben ismertetett modellt használtuk.

[1] a rövidítésekben az angol megfelelőt használtuk, pl. Manual Transcript - Manual Punctuation

[2] Esetenként ez nagy kihívás, amennyiben a szóhibák miatt az eredeti írásjelezés értelmét veszti.

2.2. Adatbázisok

Kísérleteinket angol és magyar nyelven végeztük el. **Magyar nyelvre** 10 szöveges blokkot választottunk ki egy televíziós műsorok átiratait tartalmazó adatbázisból (Tarján és mtsai, 2016); sporthíreket, időjárás-jelentéseket és híradókat vizsgáltunk meg. Ez a részkorpusz összesen 500 mondatot, így megközelítőleg 8000 szót foglal magában. A felhasznált ASR rendszer (Varga és mtsai, 2015) szóhibaarány értékeit illetően rendre 6,8%-ot, 10,1%-ot és 21,4%-ot mértünk az időjárás-jelentések, a híradók és a sporthírek esetén. Automatikus központozáshoz a (Tündik és mtsai, 2018)-féle, magyar nyelvre adaptált modellt használtuk, melynek teljesítménye F1-mértéket tekintve 60-70% kézi átiratokon, gépi átiratokon pedig 45-50%.

Angol nyelvre az IWSLT2011 adathalmazban található TED előadások átiratai közül használtunk fel 9 szöveges blokkot (Federico és mtsai, 2012). Ez a részkorpusz összesen 800 mondatot, így megközelítőleg 12000 szót foglal magában. Az ASR átiratok a (Rousseau és mtsai, 2012) cikkben bemutatott módszerrel készültek, melyeken 18,7% -os szóhibaarányt mértünk. Automatikus központozáshoz a (Tündik és mtsai, 2018)-féle angol nyelvre adaptált modellt használtuk, melynek átlagos teljesítménye F1-mértéket tekintve 60-70% kézi átiratokon, gépi átiratokon pedig 50-55%.

A magyar és angol nyelvű referencia összefoglalók készítését 3 annotátor vállalta (minden szöveges blokkhoz 3 darab, 10-12 mondat terjedelmű összefoglaló készült), így a szóhibák és a központozási hibák által keletkezett szemantikai torzításokat egy dokumentum-összefoglaló feladat keretében is meg tudtuk vizsgálni.

3. Módszerek

Cikkünkben néhány olyan megközelítést ismertetünk és értékelünk ki, amelyek a szemantikai torzítások számszerűsítésére alkalmasak. Ezen mértékek esetén két alapvető szempont jön szóba: (i) kiszámítjuk az egyes mondatpárok (ugyanazon mondat kézi és gépi átiratának) szemantikai hasonlóságát, szóbeágyazások alapján, míg (ii) a gépi átiratból és írásjelezésből adódó hibák kölcsönhatásának elemzését tartalmi összefoglalási feladaton keresztül vizsgáljuk meg. A szemantikai torzításra vonatkozó összehasonlítást így mondat- illetve dokumentumszinten is elvégezzük.

3.1. Mondatszintű hasonlóság

Első lépésként meghatározzuk a mondatvektor-reprezentációkat egy adott mondat szóvektorainak segítségével. Angol nyelvre az előtanított GloVe (Pennington és mtsai, 2014) és word2vec (Mikolov és mtsai, 2013a) szóbeágyazásokat, magyar nyelvre pedig a „Makrai-féle" szóvektorokat (Makrai, 2016) használtuk fel vizsgálatainkhoz. Megfontoltuk a modernebb, kontextuális beágyazások és karakter N-gram sorozatokkal kiterjesztett szóvektorok használatát, de ezeket végül elvetettük, mivel nem álltak rendelkezésre magyar nyelvre a vizsgálat idején, illetve

a karakter N-gramok hozzáadását korábban kontraproduktívnak találtuk, valószínűleg a magyar nyelv extrém gazdag morfológiája és kötetlen szórendje miatt. (Azt tapasztaltuk, hogy a szóvektorok szépen megtanulják a morfoszintaxist, de összességében szinte teljesen elveszítik a szemantikus konzisztenciát).

Továbbá a mondatszintű kódolók (Cer és mtsai, 2018; Conneau és mtsai, 2017) alkalmazását is mellőztük, elsősorban azért, mert az általunk ismertetett, egyszerűbb megközelítések hasonló teljesítményt mutatnak ezekkel a nehéz és összetett megközelítésekkel (Ethayarajh, 2018). Ily módon nem kellett megküzdenünk olyan nehézségekkel sem, mint például a magyar nyelvre történő adaptálás; ehelyett inkább kihasználjuk a kevésbé bonyolult, felügyeletlen megközelítések összes előnyét. A következő vektorábrázolási formákat használjuk a szemantikai torzítás/hasonlóság mondatszintű vizsgálatára:

Szózsák (Bag-of-Words, BOW): a legegyszerűbb vektorizálási formában a mondat szavainak egyszerű átlagát vesszük. Esetlegesen stop-szó szűrést végzünk az NLTK könyvtárral.

Simított Inverz Gyakoriság (Smooth Inverse Frequency, SIF): A SIF mondatbeágyazások (Arora és mtsai, 2016), súlyozottan átlagolják a szóvektorokat. A súlyokat (W) az alábbi formulával számíthatjuk:

$$W(w_i) = \frac{a}{a + p(w)}, \tag{1}$$

ahol a a simítást befolyásoló paraméter (alapértelmezetten $a = 0,001$), $p(w_i)$ pedig a w_i szó referencia korpuszon számított relatív gyakorisága. Ily módon a gyakori szavak súlya kisebb, a szemantikailag relevánsabbaké pedig nagyobb lesz. Az ezt követő lépésben a SIF vektorokat konkatenáljuk egy mátrixba, amelyet szinguláris érték felbontással (SVD) felbontunk. A SIF mondatvektorok első szinguláris értékre vett projekcióját ezután kivonjuk a súlyozott átlagból, így csökkentve a szemantikailag nem odaillő szavak befolyását.

Nem felügyelt SIF (uSIF): az uSIF (Ethayarajh, 2018) módszer az előbb bemutatott SIF reprezentációhoz képest abban különbözik, hogy a értékét is közvetlenül becsüljük a gyakoriság szerint rendezett szótárból. Az első m szinguláris értéket őrizzük meg, rendre $\lambda_1 ... \lambda_m$ súlyokkal:

$$\lambda_i = \frac{\sigma_i^2}{\sum_{j=1}^{m} \sigma_i^2}, \tag{2}$$

ahol σ_i a mondatbeágyazó mátrix i-edik szinguláris értéke. Látható, hogy $m = 1$, esetén az uSIF a SIF-fel azonos, amennyiben a-t optimalizáltnak tekintjük. m leggyakrabban választott értéke 5.

A mondatok közötti hasonlóság mérésére páronként hasonlítjuk össze az egymáshoz illesztett mondatok szekvenciáit:

$$sim(a, b) = \frac{\sum_{i=0}^{S-1} a_i b_i}{\sum_{i=0}^{S-1} a_i^2 \sum_{i=0}^{S-1} b_i^2} \tag{3}$$

ahol a és b a két mondatbeágyazó vektor(melyek származtathatóak akár a BOW, a SIF vagy az uSIF eljárással) az S dimenziós mondatbeágyazó térben.

A mondatok közötti hasonlóságot egy negyedik módon, közvetlenül a szóvektorokból is származtathatjuk: a **Word Mover's Distance** (WMD) egy népszerű módszer dokumentumok / mondatok összehasonlítására (Kusner és mtsai, 2015). Alapja, hogy az összehasonlítandó dokumentumok (vagy esetünkben mondatok) között a szemantikus térben megadja azt a legkisebb költségű utat, amellyel a két dokumentum (mondat) egymásba átvihető. A WMD a népszerű Gensim python könyvtárban is implementált. A WMD alapján a hasonlóságot egyszerűen számíthatjuk két mondat közt:

$$WMS = \frac{1}{1 + WMD}. \tag{4}$$

3.2. Dokumentumszintű hasonlóság

A gépi beszédfelismerés egyik izgalmas felhasználási területe a beszélt nyelvi dokumentumok, rekordok tartalmi kivonatolása, összefoglalása. Ennek során beszédfelismerővel átírjuk a beszédet, majd az így nyert szövegen futtatjuk a tartalmi összefoglaló algoritmust.

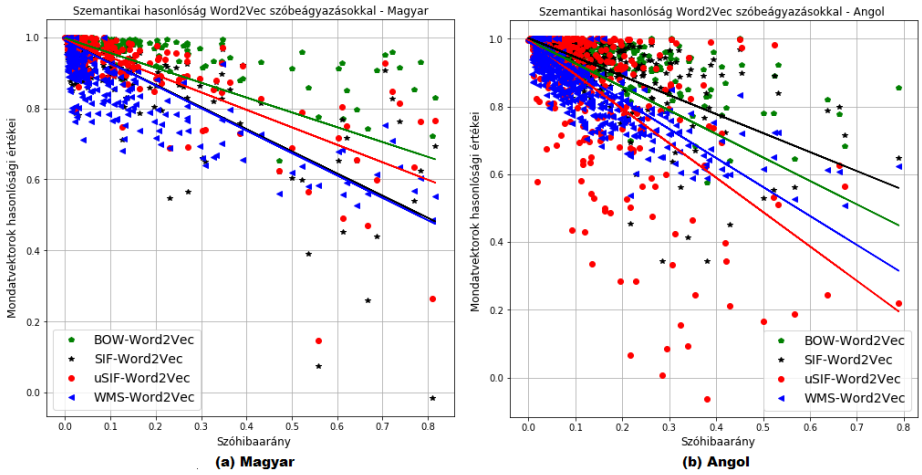
A kísérlethez az MT-MP, AT-MP, MT-AP és AT-AP eljárásokkal nyert szövegeket vesszük alapul, és valamennyire tartalmi összefoglalót generálunk. Az egyes összefoglalók közötti különbséget a Recall-Oriented Understudy for Gisting Evaluation eljárással, rövidebb nevén a ROUGE metrikákkal mérjük (Lin, 2004). A ROUGE többféle összehasonlítást is lehetővé tesz, ezek részletes ismertetése meghaladná jelen cikk kereteit, de kimerítő leírás található például a (Lin, 2004) irodalomban. Jelen munkában az alábbi ROUGE metrikákat használjuk:

- ROUGE-1: unigram (szavankénti) átfedést mér (felidézésben);
- ROUGE-2: bigram (szókettesek szerinti) fedést mér (a kérdéses összefoglaló milyen arányban idézi fel a referencia szóketteseit);
- ROUGE-L: leghosszabb közös szószekvencia;
- ROUGE-SU4: skip-bigram és N-gram alapján méri az együttes előfordulást (szinonimákat is kezeli a skip-gram révén).

Referenciaként a 3 független annotátor által az MT-MP szövegek alapján készített összefoglalókat használjuk (mivel többféle összefoglaló is készíthető, bevett gyakorlat nem egyetlen referenciával összevetni a kimenetet). A gépi tartalmi összefoglalást a Gensim modul (Mihalcea és Tarau, 2004) BM25 rangsoroló eljárásával (Barrios és mtsai, 2016) készítjük. Bár a BM25 több mint 10 éve ismert összefoglaló algoritmus, azért esett erre a választásunk, mert ipari alkalmazásokban is megtaláljuk, illetve mert nagyon egyszerűen használható, nem igényel adaptációt sem. Ugyanezen okokból mellőztük a beágyazásokon alapuló algoritmusokat is, illetve azért is, mert nem jellemző, hogy a felismerő szinonimára tévesszen, sokkal inkább hangzásában hasonló szóra. Mindazonáltal a jövőben mindenképp érdemes a kísérletet szemantikus reprezentációk alapján működő összefoglaló algoritmusokkal is elvégezni.

4. Eredmények és Diszkusszió

A mondatszintű kiértékelés esetében az MT-MP és az AT-MP átiratokat hasonlítottuk össze, mivel a kézi és az automatikus központozással készült dokumentumok mondatainak egymáshoz igazítása nem triviális feladat: az írásjelek megváltoztathatják a mondathatárokat, így a központozás típusai (MP és AP) szerinti összehasonlítás jobban illeszkedik a dokumentumszintű megközelítéshez. Az 1. ábra az MT-MP és az AT-MP átiratok mondatpárjain vett szemantikai hasonlósági értékeket (BOW, SIF, uSIF és WMS) ábrázolja, magyar (a) és angol (b) nyelvre. Így az x tengelyen lévő szóhibaarány is a mondat szintjén értendő.



1. ábra: Mondatszintű szemantikai hasonlósági értékek a szóhibaarány (WER) függvényében

Figyelembe véve a valós ASR-felhasználási eseteket, ahol a $WER < 30\%$ magyar nyelvre, angol nyelvre pedig $WER < 20\%$ értékű[3], a szemantikai térre gyakorolt hatás korlátozott, a hasonlósági értékek legtöbbje 0,8 és afölött van. Érdemes megvizsgálni a szórásokat is, melyek mértéke WER=20% felett látványos emelkedést mutat. Az MT-MP és AT-MP átiratok mondatain vett hasonlóságok átlagait az 1. táblázat mutatja, ahol a szóhibák ellenére nagyon magas szemantikai egyezést figyelhetünk meg. A magyar nyelvű kísérleteinkhez 300-dimenziós word2vec és 152-dimenziós GloVe szóbeágyazásokat használtunk. Mivel a SIF, az uSIF és a WMS kategóriák esetében a két megközelítés eredményei konzisztens trendeket mutattak, ezért csak a word2vec reprezentációkhoz tartozó eredményeket mutatjuk be.

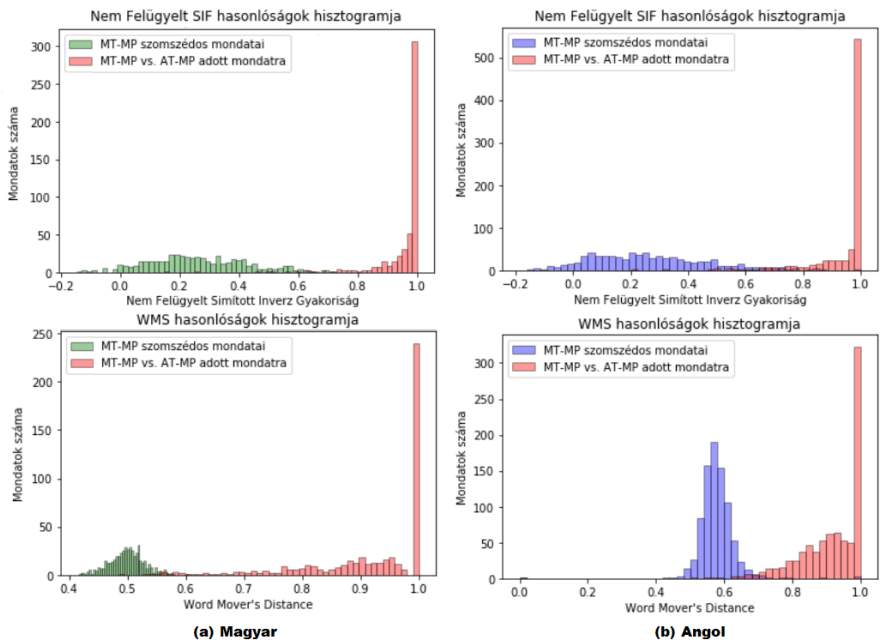
[3] A morfológiailag gazdag magyar nyelv esetén magasabb WER-érték mellett érzékeljük hasonlónak az ASR-teljesítményét (Kurimo és mtsai, 2006)

1. táblázat. Mondatszintű szemantikai hasonlósági értékek magyar és angol nyelvre

Mértékek	BOW	SIF	uSIF	WMS
Magyar	0,97	0,95	0,96	0,92
Angol	0,94	0,96	0,91	0,90

Ahogy az várható volt, nincs szignifikáns különbség a szóvektorok két típusa között. A BOW megközelítést illetően a szóvektorok két típusa kvázi-ekvivalenssé válik, amikor a mondatvektorok kiszámítása esetén egy előzetes stop-szó szűrést alkalmazunk az adott mondathoz tartozó GloVe szóvektorok átlagolásakor. Ez érthető, mivel a word2vec módszer esetén a stop-szavakat alulmintavételezik (Mikolov és mtsai, 2013b), míg a GloVe tanítása során megőrzik.

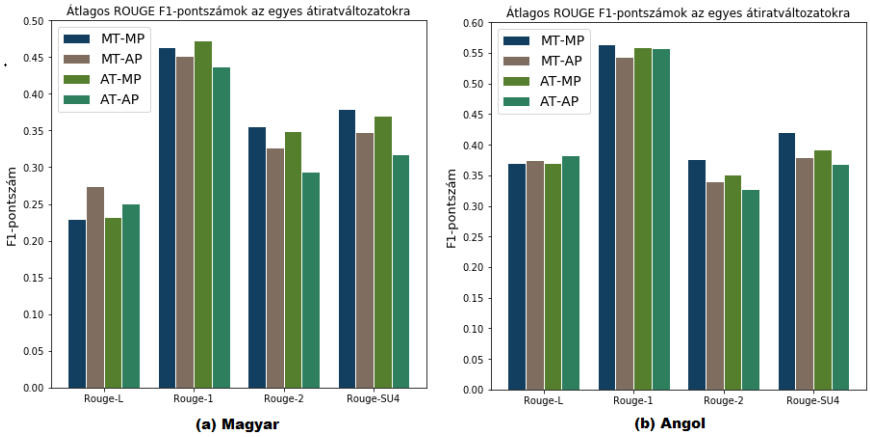
Egyfajta referenciaértékek felállítása érdekében – tekintettel az 1. ábrán látható MT-MP és AT-MP átiratváltozatok közötti hasonlósági értékre – az MT-MP típusú dokumentumban a szomszédos mondatok szemantikai hasonlóságainak eloszlását is meghatároztuk. Ennek a lépésnek az a célja, hogy össze tudjuk hasonlítani a szóhibákból származó mondatonkénti szemantikai változásokat a referenciadokumentum mondatai között megfigyelhető szemantikai hasonlósággal. A 2. ábra az uSIF és WMS mértékek eloszlását mutatja.



2. ábra: Szemantikai hasonlóságok (uSIF és WMS) eloszlásának ábrázolása hisztogrammal, szomszédos mondatok között a kézi átiratban, ill. ugyanazon mondat kézi és gépi átiratai között

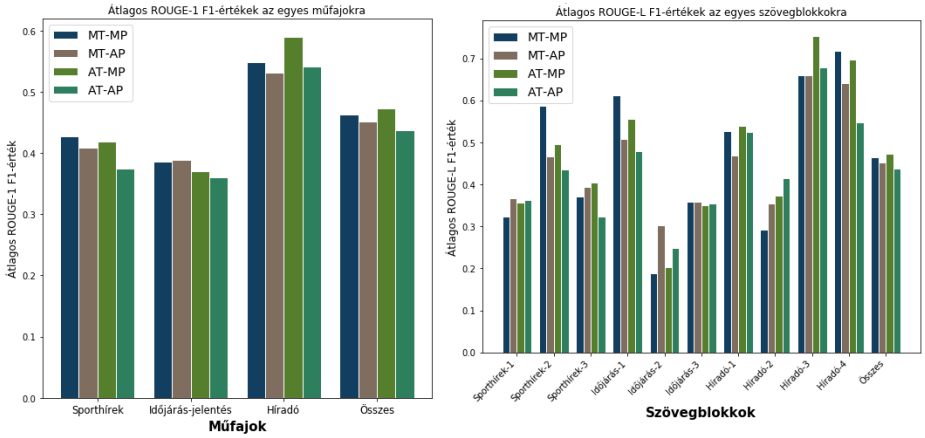
Mindegyik ábrán két hisztogram látható: a kézi és gépi átiratok közötti hasonlóságok eloszlását és a kézi átiraton belül, a szomszédos mondatok közötti hasonlóságok eloszlását. A két eloszlás között alig van átfedés, a magyar nyelv (lásd. 2. ábra 'a)' része) és az angol nyelv (lásd. 2. ábra 'b)' része) esetében. Ez azt jelenti, hogy a szóhibákból eredő szemantikai torzítás nem olyan magas, hogy egy tévesen felismert mondatot közelebb hozzon a szomszédos mondatok jelentéséhez, mint az eredeti jelentéshez. Figyelembe véve, hogy a szomszédos mondatok tipikusan közelebb állnak a szemantikai térben, mint ugyanazon dokumentumon belül a nem szomszédos mondatok, ez meglehetősen kielégítő eredmény, amely megmagyarázza azt is, hogy a tapasztalatokkal összhangban a jelentés kinyerése hogyan lehet kellőképpen robusztus azokból a mondatokból, amelyek szóhibákat tartalmaznak.

Rátérve a tartalmi összefoglalás feladatra, a 2.1. fejezetben ismertetett átiratváltozatokra vonatkozó ROUGE eredményeket a 3. ábra illusztrálja, magyar és angol nyelvre. Mivel a magyar nyelvű adatbázis különféle műfajú szövegeket tartalmazott, ezért eredményeinket a 4. ábrán műfaj szerinti bontásban, és az egyes blokkokat tekintve is bemutatjuk.



3. ábra: Tartalmi kivonatolás kiértékelése magyar és angol nyelvre

Az egyik legfontosabb szempont a „tökéletes" MT-MP és a valós felhasználást tükröző AT-AP átiratváltozatok eredményeinek összehasonlítása (utóbbinál mind az átirat, mind a központozás automatikusan történik). A különböző műfajokra vonatkozó magyar nyelvű tartalmi kivonatolási eredményeket szemlélve a 4. ábrán, az AT-MP átiratok eredménye szorosan korrelál az ASR pontossággal (sporthírek és híradók esetében), valószínűleg azért, mert az ASR rendszer nyelvi modelljének és a tartalmi kivonatoló szemantikai rangsoroló moduljának hasonló nyelvi komplexitású feladattal kell megbirkóznia. Az időjárás-jelentések kivételt képeznek; feltételezzük, hogy a gyakoriság alapú kivonatolási megközelítés kevésbé alkalmas ilyen típusú dokumentumokhoz.



4. ábra: Tartalmi kivonatolás magyar nyelvre, műfaji és blokkonkénti bontásban

A legjobb kivonatolási eredményeket a híradó kategóriájára kaptuk, annak ellenére, hogy a gépi átirat időjárás-jelentések esetében pontosabb volt. Az időjárás-jelentések esetében viszont a központozás pontatlanabb (Tündik és mtsai, 2018), a legkevésbé precíz automatikus központozás pedig a sporthírek kategóriájához társul (Tündik és mtsai, 2018).

Láthatjuk, hogy az írásjelekkel kapcsolatos hibák fontosabbak a kivonatolás szempontjából. Ez korrelál a mondatonkénti szemantikai vizsgálatainknál látottakkal: a szóhibák korlátozott torzítást eredményeznek a szemantikai térben a mondatok szintjén, feltéve, hogy a valódi mondathatárok ismertek (AT-MP). A 3. ábrán látható ROUGE-pontszámokra kitérve, a ROUGE-2 és a ROUGE-SU4 esetében megfigyelhető, hogy az MT-AP kategóriára vonatkozó értékek alacsonyabbak, mint az AT-MP esetében, valamint az, hogy az eredmények közötti különbség nagyobb, ha a központozás módját változtatjuk (kéziről automatikusra), mint amikor az átirat típusa változik (kéziről automatikusra).

Az AT-MP és az AT-AP kategóriák összefoglalóit összehasonlítva, a ROUGE-2 és a ROUGE-SU4 pontok szerinti különbség jelentős. Habár az AT-AP esetben a szóhibák már az automatikus központozásba is továbbterjednek, eredményeink azt igazolják, hogy a mondatszintű tokenizálási (központozási) hibák nagyobb mértékben befolyásolják a kivonatolást, mint a szóhibák. Az eredmények azt sugallják, hogy a mondatokra bontás esetében javallott a prozódiai jellemzőkre is támaszkodni, amelyek a szóhibákkal szemben jóval robusztusabbak, mint a szöveges jellemzők. A jövőben mind prozódiai alapú, közvetlen szegmentálási módszereket (pl. (Beke és Szaszák, 2016)), mind akusztikai-szöveges központozási megoldásokat (pl. (Szaszák és Tündik, 2019)) is érdemes megvizsgálni tartalmi kivonatoláskor.

5. Összegzés

Cikkünkben megvizsgáltuk a szóhibák és központozási hibák által kiváltott szemantikai torzítást. Az ASR rendszerekből származó szóhibák már az automatikus központozás feladatába továbbterjednek, amikor a nyers gépi átirat tokenizálása a cél; ezután pedig mindkét (szó- és központozási) hibatípussal számolni kell a tartalmi összefoglalók készítése esetében.

Egyszerű, mondatszintű hasonlósági metrikákkal bebizonyítottuk, hogy a szóhibák jelenléte kisebb mértékű torzítást eredményez a szemantikai hasonlóságban ugyanazon mondatot vizsgálva, mintha két, szomszédos mondat közötti szemantikai különbséget vizsgálnánk. Valójában a két eset hasonlósági eloszlása marginális átfedést mutatott, ami azt sugallja, hogy a szóhibák ritkán okoznak drámai eltolódást a szemantikai térben a mondatok szintjén (és ennélfogva magasabb szinteknél, pl. a dokumentumok szintjén).

Mivel a gépi átiratban elveszik a valós mondatszint, automatikus központozást kell alkalmazni. A szemantikai torzítás tartalmi összefoglalások vizsgálatának szemszögéből történő értékelése lehetővé tette számunkra, hogy elemezzük az írásjelhibákat is a szóhibák mellett. Megállapítottuk, hogy az írásjelek miatt a ROUGE-2 és a ROUGE-SU4 pontszámok közötti relatív különbség nagyobb, mint a szóhibák esetén, bár a szóhibák az írásjelezési feladatra is hatást gyakorolnak. A teljesen automatikus (AT-AP) bemenetű összefoglalók elemzése azonban azt mutatta, hogy az ASR-feldolgozási lánc jelenlegi szűk keresztmetszete elsősorban a központozás okozta mondatszintű eltérésből fakad, nem pedig a szóhibákból, még a szóhibaarány 20%-hoz közeli szintjén is. Ezek a megállapítások extraktív tartalmi összefoglalásra érvényesek, absztraktív változat vizsgálatára a magyar nyelv korlátai miatt nem nyílt lehetőségünk.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely az FK-124413 projekt keretében a cikkben ismertetésre került kutatást támogatta. Köszönjük továbbá az NVIDIA támogatását (GPU biztosítása a neurális hálózatok tanításához).

Hivatkozások

Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2016)

Barrios, F., López, F., Argerich, L., Wachenchauzer, R.: Variations of the similarity function of textrank for automated summarization. arXiv preprint arXiv:1602.03606 (2016)

Beke, A., Szaszák, G.: Automatic summarization of highly spontaneous speech. In: International Conference on Speech and Computer. pp. 140–147. Springer (2016)

Celikyilmaz, A., Hakkani-Tür, D.: Discovery of topically coherent sentences for extractive summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 491–499. Association for Computational Linguistics (2011)

Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., és mtsai: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)

Ethayarajh, K.: Unsupervised random walk sentence embeddings: A strong but simple baseline. In: Proceedings of The Third Workshop on Representation Learning for NLP. pp. 91–100 (2018)

Federico, M., Stüker, S., Bentivogli, L., Paul, M., Cettolo, M., Herrmann, T., Niehues, J., Moretti, G.: The IWSLT 2011 evaluation campaign on automatic talk translation. In: International Conference on Language Resources and Evaluation (LREC). pp. 3543–3550 (2012)

Genest, P.E., Lapalme, G.: Fully abstractive approach to guided summarization. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 354–358 (2012)

Kafle, S., Huenerfauth, M.: Effect of speech recognition errors on text understandability for people who are deaf or hard of hearing. In: Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT). pp. 20–25 (2016)

Klejch, O., Bell, P., Renals, S.: Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5700–5704. IEEE (2017)

Kröger, B.J., Crawford, E., Bekolay, T., Eliasmith, C.: Modeling interactions between speech production and perception: speech error detection at semantic and phonological levels and the inner speech loop. Frontiers in Computational Neuroscience 10, 51 (2016)

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., Saraclar, M.: Unlimited vocabulary speech recognition for agglutinative languages. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 487–494. Association for Computational Linguistics (2006)

Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning. pp. 957–966 (2015)

Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)

Makrai, M.: Filtering Wiktionary triangles by linear mapping between distributed models. In: Proceedings of LREC. pp. 2776–2770 (2016)

Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–411 (2004)

Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013a)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (szerk.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013b)

Öktem, A., Farrús, M., Wanner, L.: Attentional parallel RNNs for generating punctuation in transcribed speech. In: International Conference on Statistical Language and Speech Processing. pp. 131–142. Springer (2017)

Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304 (2017)

Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of EMNLP. pp. 1532–1543 (2014)

Postma, A.: Detection of errors during speech production: A review of speech monitoring models. Cognition 77(2), 97–132 (2000)

Rousseau, A., Deléglise, P., Esteve, Y.: TED-LIUM: An automatic speech recognition dedicated corpus. In: LREC. pp. 125–129 (2012)

Simonnet, E., Ghannay, S., Camelin, N., Estève, Y.: Simulating ASR errors for training SLU systems. In: LREC 2018 (2018)

Szaszák, G., Tündik, M.Á.: Leveraging a character, word and prosody triplet for an ASR error robust and agglutination friendly punctuation approach. Proc. Interspeech 2019 pp. 2988–2992 (2019)

Tarján, B., Varga, Á., Tobler, Z., Szaszák, Gy., Fegyó, T., Bordás, Cs., Mihajlik, P.: Magyar nyelvű, élő közéleti- és hírműsorok gépi feliratozása. In: XII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2016. pp. 89–99. Szeged (2016)

Tündik, M.Á., Szaszák, G.: Joint word- and character-level embedding CNN-RNN models for punctuation restoration. In: 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000135–000140. IEEE (2018)

Tündik, M.A., Szaszák, G., Gosztolya, G., Beke, A.: User-centric evaluation of automatic punctuation in ASR closed captioning. In: Proc. Interspeech 2018. pp. 2628–2632 (2018)

Varga, Á., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T., Bordás, C., Mihajlik, P.: Automatic close captioning for live Hungarian television broadcast speech: A fast and resource-efficient approach. In: International Conference on Speech and Computer. pp. 105–112. Springer (2015)

Voleti, R., Liss, J.M., Berisha, V.: Investigating the effects of word substitution errors on sentence embeddings. arXiv preprint arXiv:1811.07021 (2018)

Poszter, laptopos bemutató II.

apPILkáció: egy Android-alkalmazás manysi nyelvtanulás céljára

Bobály Gábor[1], Horváth Csilla[2,3], Vincze Veronika[4]

[1]IT Services Hungary
[2]MTA Nyelvtudományi Intézet
[3]University of Helsinki
[4]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
{bobalygabor,naj.agi}@gmail.com,vinczev@inf.u-szeged.hu

Kivonat A cikkben bemutatjuk Android-alkalmazásunkat, mely a manysi nyelv oktatását, konkrétabban manysi szavak tanítását célozza. Az alkalmazás nyelve választhatóan magyar, orosz vagy angol. Moduljai között találunk általános szókincsre épülő, véletlenszerűen kiválasztott manysi szavakat tanító modult, illetve meghatározott szemantikai mezőhöz (pl. állatok) tartozó szavakra épülő szókitaláló modult. Az alkalmazás elsődlegesen a manysi nyelvtanulók érdeklődésére tarthat számot, de hasznos lehet nyelvészeknek vagy nyelvtanároknak is. Az alkalmazást mindenki számára ingyenesen elérhetővé tesszük.
Kulcsszavak: Android, manysi, nyelvtanulás, veszélyeztetett nyelv

1. Bevezetés

Manapság az internet és a digitális technológiák elterjedése számos lehetőséget kínál a valós idejű kommunikációra a világ bármely részén élő emberek között. Ezt elősegítik a különféle nyelvtechnológiai eszközök, mint például a beszédfelismerő rendszerek, helyesírás-ellenőrzők és gépi fordítórendszerek, hogy csak néhányat említsünk a sok közül. Azonban a kisebbségi nyelvek esetében legtöbb esetben nem állnak rendelkezésre ezek az eszközök, ami hosszú távon e nyelvek digitális térben való használatát igencsak leszűkíti. Meg kell említenünk azonban, hogy a világban számos projekt célozza a veszélyeztetett nyelvek revitalizációját a digitális térben: ezek fő célja különféle digitális eszközök és erőforrások előállítása a szóban forgó nyelvekre.

Jelen cikk egy Nyugat-Szibériában beszélt veszélyeztetett őshonos nyelvre, a manysira összpontosít. Ugyan a manysi beszélők száma csökken, emelkedik a nyelvismeret és a nyelvhasználat presztízse, valamint nő a nyelvelsajátítás és örökségnyelvi nyelvelsajátítás iránti érdeklődés is, különös tekintettel a városi manysi nyelvtanulókra. A manysi nyelvtanulási törekvéseket elősegítendő, megalkottunk egy Android-alkalmazást, melyet bárki számára szabadon elérhetővé teszünk. E cikkben bemutatjuk az apPILkáció nevű alkalmazást, főbb funkcióival együtt. Ismereteink szerint ez az első olyan okostelefonos alkalmazás, mely kifejezetten a manysi nyelvtanulás támogatására hivatott.

A cikk felépítése a következő. Először áttekintjük a jelenleg is létező főbb nyelvtanító mobilos alkalmazásokat, majd röviden ismertetjük a manysi nyelvet beszélők és tanulók helyzetét. Ezután bemutatjuk az apPILkációt a mögöttes szótárakkal és a főbb modulokkal együtt. A cikket az apPILkáció alkalmazási lehetőségeinek felvázolásával zárjuk.

2. Háttér

A mobiltelefonnal támogatott nyelvtanulás (mobile-assisted language learning – MALL) a nyelvoktatás gyorsan fejlődő ága az okostelefonok egyre nagyobb elterjedtségének köszönhetően világszerte (Chinnery, 2006). Traxler (2005) definíciója szerint a mobilalapú tanulás olyan oktatási módszertan, ahol az egyetlen vagy túlnyomórészt használatos technológiák kézben tartható eszközökre épülnek. A MALL lehetővé teszi, hogy a tanulók bárhol, bármilyen körülmények között tudjanak tanulni, az asztali számítógépek kötöttségeitől mentesen (Miangah és Nezarat, 2012). Ennek ellenére a MALL-t kutató tanulmányok legtöbbje csak az intézményes mobilalapú oktatásra fókuszál, viszonylag kevés figyelmet szentelve az osztálytermen kívüli használatra (Godwin-Jones, 2017). Például, Stockwell és Hubbard (2013) tíz alapvető elvet fogalmaz meg a MALL-lal kapcsolatosan, hangsúlyozva többek közt a nyelvtanulók közti egyéni különbségekre való odafigyelést és a feladatok rövidségének fontosságát.

Számos nyelvoktató okostelefonos alkalmazás létezik a világban, a kétnyelvű szótáraktól kezdve a nyelvtani gyakorlatokat is tartalmazó eszközökig. Az ily módon tanulható nyelvek száma azonban korlátozott, főként a nagyobb nyelvekre léteznek megoldások. A teljesség igénye nélkül, okostelefonos nyelvi kurzusokat kínáló alkalmazások például a Babbel[1], a Duolingo[2], a Memrise[3] és a Busuu[4]. Az 1. táblázat mutatja a fenti alkalmazásokban jelenleg (2019 novembere) tanulható nyelveket. Az anyanyelvi beszélőkre vonatkozó adatok az angol Wikipediából származnak.

Láthatjuk a táblázatból, hogy elsősorban világnyelveket, illetve a kisebb nyelvek közül többségében Európában beszélt nyelveket tanulhatunk az alkalmazásokban, nem is beszélve a holt nyelvekről (latin) vagy a kitalált nyelvekről (az eszperantón felül klingonul és nemes valírul is tanulhatunk a Duolingo segítségével, itt két utóbbi nyelv egy-egy tévésorozatnak köszönheti népszerűségét).

Ami a kisebbségi uráli nyelvek mobiltelefonnal támogatott oktatását illeti, csak néhány alkalmazásról van tudomásunk, például a Laring[5], amelyet a Tromsøi Egyetemen fejlesztettek ki a déli számi oktatására.[6] Különböző szócsoportokba tartozó szavakat tanít a rendszer, hangosan felolvassa a norvég szó déli

[1] https://www.babbel.com/
[2] https://www.duolingo.com/
[3] https://www.memrise.com/
[4] https://www.busuu.com/
[5] http://divvun.no/laring/laring.html
[6] A cikk írásakor csak az iPhone-os változatot tudtuk tesztelni, az androidos változat sajnos nem bizonyult elérhetőnek.

Nyelv	Anyanyelvi beszélők száma	Babbel	Duolingo	Memrise	Busuu
kínai	1500M		•	•	•
spanyol	400M	•	•	•	•
angol	332M		•		•
hindi	370M		•		
arab	300M		•	•	•
portugál	230M	•	•	•	•
francia	220M	•	•	•	•
orosz	145M	•	•	•	•
japán	126M		•	•	•
német	90M	•	•	•	•
koreai	78M		•	•	
vietnami	70M		•		
olasz	63M	•	•	•	•
török	60M	•	•	•	•
lengyel	50M	•	•	•	•
indonéz	43M	•	•		
ukrán	35M		•		
román	24M		•		
holland	22M	•	•	•	
görög	20M		•		
magyar	15M		•		
cseh	12M		•		
katalán	10M		•		
svéd	9M	•	•	•	
héber	6M		•		
dán	5,5M	•	•	•	
guarani	4,8M		•		
norvég	4,6M	•	•	•	
mongol	3,6M			•	
szlovén	2,5M			•	
szuahéli	2M		•		
walesi	610K		•		
izlandi	310K			•	
ír	260K		•		
navahó	170K		•		
hawaii	2K		•		
eszperantó	0		•		
klingon	0		•		
latin	0		•		
nemes valír	0		•		

1. táblázat. Okostelefonos alkalmazásokban tanított nyelvek.

számi megfelelőjét, illetve egy másik feladatban a hallás utáni értést is lehet gyakorolni: ki kell választani a hallott számi szót jelölő képet.

Mindemellett a Memrise felhasználók által létrehozott kurzusai között is találunk uráli, ileltve szibériai nyelvekre létrehozott anyagokat, többek között az alábbi nyelvekre: inkeri[7], lív[8], kvén[9] és jakut[10]. Előfordul, hogy e kurzusok csak néhány tucat szót ölelnek fel, mint például az ulcs[11] és az enyec[12] esetében. A manysi kurzus mindösszesen 11 szót tartalmaz, ebből egy szó helytelen fordítással szerepel, valamint a szavak helyesírása se következetes. Az apPILkáció szóállománya ennél jóval több, továbbá a helyesírás is a jelenlegi normákat követi, így minden esély megvan arra, hogy nagyobb érdeklődést váltson ki a potenciális nyelvtanulók körében, mint a fenti alkalmazás.

3. Manysi beszélők és nyelvtanulók

A manysi egy Nyugat-Szibériában beszélt veszélyeztetett nyelv. A dominánsan orosz nyelvű, többnemzetiségű és többnyelvű közegben marginális szerepet tölt be, használatát a hagyományosnak tekintett életmód visszaszorulása és a gyors urbanizálódás is hátráltatja. Ugyan a manysikat hagyományosan félnomád életmódot folytató közösségnek tekintették (és bizonyos mértékben tekintik ma is), napjainkban a manysik többsége többnemzetiségű városokban él.

A manysi irodalmi nyelv és az írott sztenderd nyelvváltozat formálását befolyásoló szovjet nyelvpolitikai tendenciák időről időre változtak. Az első, latin betűs manysi ábécét 1931-ben hozták létre az Északi Népek Intézetében. Rövid ideig volt használatban, 1937-ben a manysi nyelvtervezők is cirill betűs írásmód használatára kényszerültek. A manysi írásbeliség ettől kezdve a cirill ábécén alapul, mindössze kisebb változásokon ment keresztül. A magánhangók hosszúságának, illetve az orosz ábécéből hiányzó speciális karaktereknek nyomtatásban való jelölése, megjelenítése az 1980-as évektől kezdve figyelhető meg. Az 1990-es évek óta két párhuzamos, mindössze egy fonéma lejegyzésében különböző írásmód van használatban, egyik a specialisták (jellemzően nyelvészek) által publikált, kisebb mennyiségű szövegben, másik az újságírók munkájának eredményeként megjelent nagy mennyiségű, a manysi nyelvhasználók szélesebb körét elérő szövegekben. A manysi nyelv történetét és státuszát figyelembe véve alkalmazásunk a szélesebb körben elterjedt írott irodalmi manysi sztenderdet, vagyis a manysi sajtóban olvasható cirill betűs írásmódot használja.

Bár a manysi nyelv és kultúra presztízse nő, a beszélők száma kritikusan alacsony. A manysi beszélőket hagyományosan három korcsoportba szokás sorolni (cf. Skribnik és Koshkaryova (2006)). A legidősebb beszélők egynyelvű manysi településen élő manysi családokban születtek és nőttek fel, túlnyomórészt maguk

[7] https://decks.memrise.com/course/2107565/ingrian/
[8] https://decks.memrise.com/course/5603933/livonian/
[9] https://decks.memrise.com/course/5596403/kven/
[10] https://decks.memrise.com/course/362501/basic-yakut/
[11] https://decks.memrise.com/course/1064732/ulch-language/
[12] https://decks.memrise.com/course/1843983/family-words-in-enets/

is manysi egynyelvűek maradtak, korlátozott orosz nyelvtudással. A középkorú beszélők szintén egynyelvű manysi családokban születtek és nőttek föl, a manysit anyanyelvükként beszélik, mindazonáltal az oktatási rendszernek köszönhetően kiegyensúlyozott orosz-manysi kétnyelvűekké váltak, és többségük dominánsan orosz nyelvű, többnemzetiségű településen él. A legfiatalabb beszélői generáció lényegesen kevesebb tagból áll, mint a már említettek, a Hanti-Manysi Autonóm Körzet peremén elhelyezkedő falvakban felnőtt maroknyi beszélő kivételével tulajdonképpen senki nem tekinthető közülük manysi egynyelvűnek, még a tanulmányai megkezdését megelőző rövid időszak tartamára sem. A nyelvtudás mértéke így általában szoros összefüggést mutat a beszélő életkorával: minél idősebb a beszélő, annál valószínűbb, hogy anyanyelvi szintű nyelvtudással rendelkezik. Ezt a tendenciát árnyalhatja a beszélő születési helye és lakhelye: gyakran a kisebb, manysi nyelvű településeken született és felnőtt fiatalabb beszélők is jól ismerik a nyelvet.

A manysi gyerekek többsége a manysi anyanyelvként való elsajátítását lehetővé tevő településeken kívül születik. Jellemzően többnemzetiségű, multikulturális városokban élnek, családjukban a mindennapok során az orosz nyelvhasználat a domináns. Manysi szülőjük jellemzően nem tette lehetővé számukra a manysi nyelv elsajátítását, és azt a gyermekek a szülők beszélgetését hallgatva sem tudták megtanulni (mivel a szülők többnyire egymás között is az oroszt használják), így az ilyen gyermekek számára a nyelvelsajátítás egyetlen lehetséges színtere az oktatás maradt. Néhány, jelentősebb manysi lakossággal rendelkező nagyvárosban alternatív oktatási intézmények működnek, melyek manysi kulturális és nyelvi ismeretek elsajátítását teszik lehetővé az olyan gyermekek számára, akik ezt a tudást nem tudták családjukban megszerezni. Ugyanakkor ezek az intézmények sem szolgálnak a nyelvhasználat stabil városi színteréül (cf. Horváth (2015, 2016)).

A manysi nyelvet (is) használó családi háttérrel vagy egyéb nyelvi környezettel kapcsolatba nem került, vagy a manysi nyelvet egyéb okból el nem sajátított, középkorú, és mindenekelőtt a fiatal nyelvtanulók, szervezett vagy egyéni nyelvelsajátítás iránt érdeklődők csoportja képezi elsődleges célcsoportunkat.

4. apPILkáció

Android-alkalmazásunkat apPILkációnak kereszteltük el, angol neve apPILcation, orosz neve PILozhenie, mivel a *pil* szó jelentése „bogyó” a manysi nyelvben. A névadással arra törekedtünk, hogy a nyelvtanulók úgy érezzék, olyan könnyű felszedegetni a szavakat az alkalmazás segítségével, mint bogyót gyűjteni az erdőben, utalva ezzel a manysi tradíciókra.

4.1. Az alkalmazás szótára

Körülbelül száz évvel ezelőtt a manysi nyelvet kutató nyelvészek már összeállították a nyelv szótárait, azonban ezek csak nemrég láttak napvilágot nyomtatásban

(Munkácsi és Kálmán, 1986; Kannisto, 2013). E szótárok a nyelv minden dialektusából tartalmaznak szavakat, ám sajnos mára több nyelvjárás is kihalt ezek közül. Az északi manysira elérhető több modern szótár is (Rombandeeva, 2005; Rombandeeva és Kuzakova, 1982), e nyelvjárás képezi alkalmazásunk szótárának alapját.

Alkalmazásunk szóállományát egy online manysi szótárból merítettük (Horváth és mtsai, 2017). A manysi szavak több manysi–orosz szótár PDF-változatából (Rombandeeva, 2005; Rombandeeva és Kuzakova, 1982) származnak, optikai karakterfelismerést használva, majd a különböző szótárak szóanyaga egységesítve lett. A manysi lexémák orosz fordításai szintén a fenti szótárakból származnak, míg a magyar és angol megfelelőket nyelvész szakértők fordították le[13]. Így tehát az apPILkáció felhasználói az oroszt, a magyart és az angolt is választhatják forrásnyelvként a manysi nyelv tanulásához.

A 2. táblázat az egyes nyelvekhez tartozó lexémák számát jelzi.

Nyelv	**Lexéma**
Manysi	13 948
Orosz	14 344
Magyar	2 334
Angol	458

2. táblázat. Az egyes nyelvekhez tartozó szószám a szótárban.

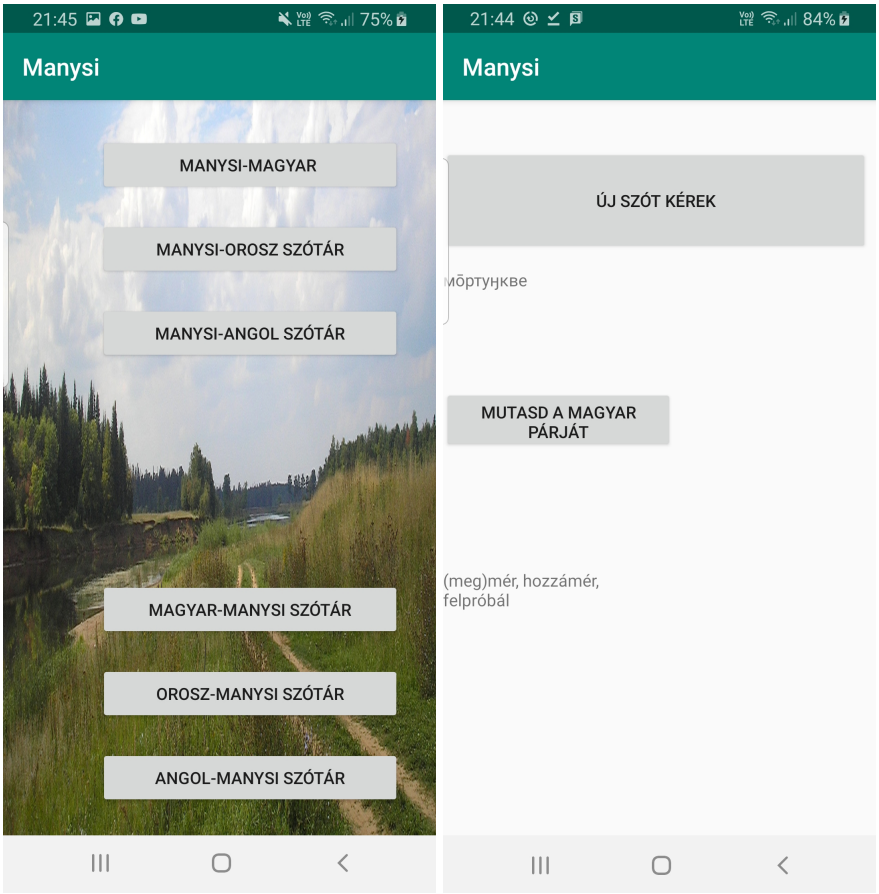
4.2. Az alkalmazás moduljai

Az apPILkáció főbb moduljai a következők. Az első modulban egy véletlenszerűen kiválasztott manysi szó jelenik meg a képernyőn, majd egy újabb gombnyomásra megjelenik a jelentése, így a felhasználó ellenőrizni tudja, hogy valóban ismeri-e a szót. A második modulban szókitalálós játékokra nyílik lehetőség, ahol a felhasználó választhatja meg a szavak tematikáját (például színek, rokonsági megnevezések, bogyós gyümölcsök vagy állatok), és a képernyőn megjelenő szónak kell kiválasztani a jelentését az alternatívák közül. Így az azonos szemantikai mezőbe tartozó szavakat együtt lehet megtanulni, illetve gyakorolni. A harmadik modulban a manysi nyelvtanról, kultúráról, illetve Manysiföld földrajzáról találhatunk alapvető információkat, mélyítve ezzel a manysi kulturális ismereteket is.

Általános szótanulás Első lépésben a felhasználó kiválasztja, melyik nyelvpárral szeretne foglalkozni: manysi–magyar, manysi–orosz, manysi–angol, magyar–manysi, orosz–manysi vagy angol–manysi (lásd az 1. ábrán). Ezután az *Új szót*

[13] A magyar és angol szóanyag jelenleg is bővítés alatt áll.

kérek gomb megérintése után megjelenik egy forrásnyelvi szó. A célnyelvi jelentést a *Mutasd a célnyelvi párját* gombra kattintva érhetjük el, lehetőséget adva a felhasználónak, hogy ellenőrizze tudását, amint az 1. ábra mutatja.

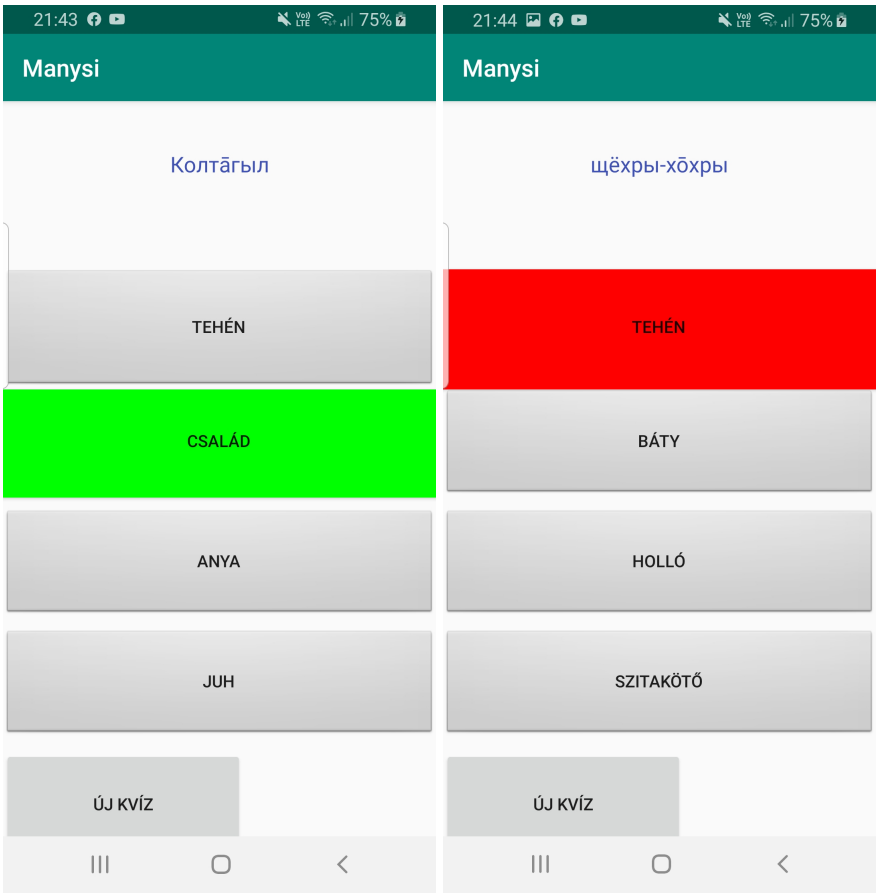


1. ábra: A választható szótárak és egy manysi–magyar szópár.

Tematikus szókitaláló játék A tematikus szókitaláló játékokban először a felhasználó kiválaszthatja, mely szemantikai mezőhöz tartozó szavakat szeretné gyakorolni. Ezt követően az adott témába sorolt szavak közül megjelenik egy véletlenszerűen kiválasztott szó, négy lehetséges célnyelvi jelentéssel: ezek egyike helyes, míg a másik három helytelen. A helyes jelentést kiválasztva a szó zöldre színeződik, jelezve, hogy ez a megfelelő jelentés (lásd a 2. ábrán). A hibás szó

választását piros szín jelzi (lásd a 3. ábrán). Az *Új kvíz* gombra kattintva új szó jelenik meg, ismét négy lehetséges jelentéssel.

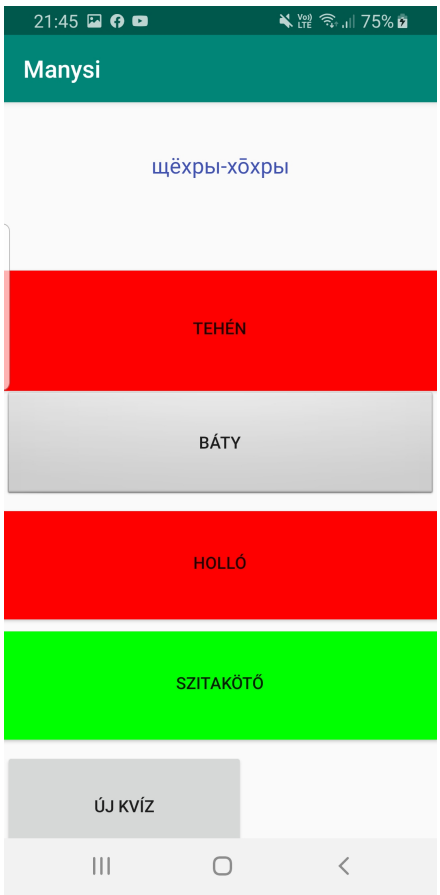
Az apPILkáció tematikus szókitalálója jelenleg négy szemantikai mezőbe tartozó szavakat tartalmaz: színek, rokonsági megnevezések, bogyós gyümölcsök és állatok. A jövőben más szemantikai csoportokkal is bővítjük az alkalmazást.



2. ábra: Egy helyes kiválasztott szó és egy helytelenül kiválasztott szó.

5. Alkalmazási lehetőségek

Az apPILkáció jelen állapotában több feladatot is képes ellátni. A fejlesztés olyan felhasználók számára ideális, akik tudnak oroszul, angolul vagy magyarul, illetve



3. ábra: Helytelen választás utáni helyes választás.

képesek a cirill betűs manysit is olvasni. Elsősorban obi-ugor nyelvekre specializálódott szakértők, kutatók és egyetemi hallgatók figyelmére számítunk, mindenekelőtt pedig a Hanti-Manysi Autonóm Körzet területén élő, manysi nyelvet tanuló diákok, hallgatók, illetve manysi nyelvtanárok érdeklődésére.

Az alkalmazás szópár modulját kezdő nyelvtanulóknak, különösen 1-5. osztályos diákoknak szánjuk. Az alkalmazás béta verziójában négy témakör szerepel, ezek száma a felhasználói visszajelzések tükrében tovább bővíthető.

A randomizált manysi szótanuló modul haladó nyelvtanulók számára bizonyulhat hasznosabbnak, elsősorban olyan felső tagozatos diákoknak és hallgatóknak, akik rendszeresen, de rövid ideig használják az alkalmazást.

Mindkét modul ideális önálló nyelvtanuláshoz. A szópár modul tanári segítséggel végzett nyelvtanulás kiegészítésére is alkalmas.

Az apPILkáció szabad hozzáférésű, ingyenesen elérhető alkalmazás lesz. Bevezetése a fejlesztőknek egyrészt a potenciálisan érdeklődő európai szakemberekkel, másrészt az offline és online manysi beszélői csoportokkal kialakított kapcsolata miatt problémamentesnek ígérkezik. Az alkalmazást a legnépszerűbb orosz közösségi portál megfelelő tematikus oldalain és közösségi chat-felületein tervezzük reklámozni, emellett beszámolókkal vagy hirdetésekkel népszerűsíthetjük az egyetlen manysi folyóiratban és az egyetlen manysi gyermekújságban is. Visszajelzésre a közösségi portálokon létrehozott oldalainkon és az alkalmazás e-mail címén keresztül számítunk, míg a sajtómunkásoktól és oktatási szakemberektől esetlegesen érkező indirekt visszajelzések manysi közvetítőinken keresztül juthatnak el hozzánk.

6. Összegzés

Ebben a tanulmányban bemutattuk az apPILkáció névre hallgató Android-alkalmazásunkat, mely manysi szavak tanítását célozza. Az alkalmazás nyelve választhatóan magyar, orosz vagy angol. Moduljai között találunk általános szókincsre épülő, véletlenszerűen kiválasztott manysi szavakat tanító modult, illetve meghatározott szemantikai mezőhöz (pl. színek) tartozó szavakra épülő szókitaláló modult. Az alkalmazás elsődlegesen a manysi nyelvtanulók érdeklődésére tarthat számot, de hasznos lehet nyelvészeknek vagy nyelvtanároknak is.

Az apPILkációt hamarosan mindenki számára ingyenesen elérhetővé tesszük.

A jövőben szeretnénk az apPILkáció szóállományát bővíteni, illetve további modulokat fejleszteni a játékos szótanulás elősegítésére. Távlati terveink között szerepel a nyelvtani gyakorlatok beépítése is az alkalmazásba. Végül szeretnénk létrehozni az apPILkáció iPhone-os változatát, az ALMAlkazást (APPLEcation / priLOMTzhenie).

Köszönetnyilvánítás

A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM).

Hivatkozások

Chinnery, G.M.: Going to the MALL: Mobile Assisted Language Learning. Language Learning & Technology 10(1), 9–16 (2006)

Godwin-Jones, R.: Smartphones and language learning. Language Learning & Technology 21(2), 3–17 (2017)

Horváth, Cs.: Beading and language class. Introducing the Lylyng Soyum Children Education Centre's attempt to revitalise Ob-Ugric languages and cultures. Zeszyty Łużyckie 48, 115–127 (2015)

Horváth, Cs.: A manysi örökségnyelv oktatási kísérletei és eredményei. Általános Nyelvészeti Tanulmányok 28, 295–306 (2016)

Horváth, Cs., Szilágyi, N., Nagy, A., Vincze, V.: Language technology resources and tools for Mansi: an overview. In: Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages. St. Petersburg, Russia (January 2017)

Kannisto, A.: Wogulisches Wörterbuch. Kotimaisten Kielten Keskuksen Julkaisuja, Helsinki (2013)

Miangah, T.M., Nezarat, A.: Mobile-Assisted Language Learning. International Journal of Distributed and Parallel Systems 3(1), 309–319 (2012)

Munkácsi, B., Kálmán, B.: Wogulisches Wörterbuch. Akadémiai Kiadó, Budapest (1986)

Rombandeeva, E.I.: Russko-mansijskij slovar'. Mirall, Sankt-Peterburg (2005)

Rombandeeva, E.I., Kuzakova, E.A.: Slovar' mansijsko-russkij i russko-mansijskij. Prosvešenije, Leningrad (1982)

Skribnik, E., Koshkaryova, N.: Khanty and Mansi: the contemporary linguistic situation. In: Pentikäinen, J. (szerk.) Shamanism and northern ecology. pp. 207–218. Mouton de Gruyter, The Hague (2006)

Stockwell, G., Hubbard, P.: Some emerging principles for mobile-assisted language learning. Tech. rep., The International Research Foundation for English Language Education, Monterey, CA (2013)

Traxler, J.: Defining Mobile Learning. In: IADIS International Conference on Mobile Learning. pp. 261–266. Malta (2005)

Tárgyas szerkezetek elemzése tenzorfelbontással – áttekintő cikk

Makrai Márton

MTA Nyelvtudományi Intézet
makrai.marton@nytud.hu

Kivonat Áttekintjük a tenzorfelbontás számítógépes nyelvészeti alkalmazásait, különösen az igei argumentumstruktúrára vonatkozókat, és olyan asszociációs mértékekre hívjuk fel a figyelmet, amelyeket eddig nem használtak erre a feladatra.
Kulcsszavak: igei többértelműség, tenzorfelbontás, függőségi elemzés

1. Bevezetés

A *tenzorok* (>2-dimenziós tömbök) a mátrixok általánosításai: ahogy a mátrixok két tengely (sorok és oszlopok) mentén elrendezve tartalmaznak számokat, a tenzoroknak több *tengelyük* (más szóval *módjuk*[1]) van. Az együtteselőfordulás-mátrix szingulárisértékfelbontása (*singular value decomposition*, SVD) természetes eszközt kínál arra, hogy általánosításokat modellezzünk két mód között a kölcsönhatásokra vonatkozóan. A két módot alkothatják szavak és a dokumentumok (látens szemantikai elemzés, *latent semantic analysis*, LSA, Landauer és Dumais (1997)), szavak és függőségi kontextusaik (Levy és Goldberg, 2014a), vagy egyszerűen a cél- (avagy fókusz-) és a kontextusszavak (szokásos szóbeágyazások, Mikolov és mtsai (2013b); Levy és Goldberg (2014b); Pennington és mtsai (2014)). Turney és Pantel (2010) szerint négyféleképpen értelmezhetjük az SVD célját: mint valamiféle látens jelentés modellezését, mint zajcsökkentést, mint közvetett (avagy magasabb rendű) együttes előfordulások modellezését (vagyis amikor két szó *hasonló kontextusokban* jelenik meg), vagy mint a ritkaság csökkentését. A nyelvben az intuíciónk szerint vannak többedrendű kölcsönhatások: a *lemezjátszó szupermenesdit játszik* kifejezés furcsa (a példa Van de Cruys (2009)-ének módosítása), jóllehet azok a másodrendű kapcsolatok, hogy ⟨játszik, SUBJ, lemezjátszó⟩ és hogy ⟨játszik, OBJ, szupermenesdi⟩ tökéletesek. A mátrixfelbontás tenzorokra való általánosításai (Kolda és Bader, 2009) az ilyen háromirányú kölcsönhatások elemzéséhez nyitnak utat.

A tenzorfelbontás a neurális hálókban szereplő szóbeágyazáshoz hasonló beágyazásvektorokat biztosít minden módhoz – a mi esetünkben az alany szerepét betöltő főnevekhez, az igékhez és a tárgy szerepét betöltő főnevekhez. Annak a projektnek,

[1] Módokról különösen azokban az alkalmazásokban beszélünk, ahol különböző modalitásból származó adatokat fuzionálnak, ahogy pl. téri és idői koordinátákat az agyi képalkotásban.

amibe ez a cikk illeszkedik, az a hosszú távú motivációja, hogy szemantikai igeosztályokat nyerjünk az ige-beágyazásvektorok klaszterezésével (felügyeletlen, vagyis annotált adatot nem használó csoportosításával). Ha a klaszterek igeosztályoknak (Levin, 1993) felelnek meg[2], akkor arra számítunk, hogy a többértelmű igék, mint a fenti *játszik*, kiugrónak *(outlier)* fognak bizonyulni, hiszen a különböző használataik különböző klaszterekbe kívánkoznak.

Az utóbbi évtizedben a vektoros szómodellek (amelyek neurális hálók szóbeágyazásaiként lettek különösen ismertek (Mikolov és mtsai, 2013a)) és a tenzorfelbontási algoritmusok is figyelemre méltó mértékben fejlődtek, és nyelvtechnológiai tesztalmazokat is használtak élvonalbeli, skálázható, zajtűrő tenzorfelbontó algoritmusok tesztelésénél (Sharan és Valiant, 2017; Bailey és mtsai, 2018; Frandsen és Ge, 2019). A szótöbbértelműség – és különösen az igei szelekció valamint argumentumszerkezet – adatközpontú megértése azonban még nem mondható érettnek. Cikkünk ezt a területet igyekszik bemutatni.

Az 2. szakasz a tenzorszámításról ad egy minimális bevezetőt, és bemutat különféle asszociációs mértékeket, olyanokat is, amelyeket tudomásunk szerint még nem használtak tenzorfelbontásban. A 3. szakasz áttekinti a tenzoros nyelvészeti munkákat, különös tekintettel a bennük alkalmazott minőségi és számszerű kiértékelésre és a kapcsolódó magyar cikkekre.

2. Tenzorfelbontás

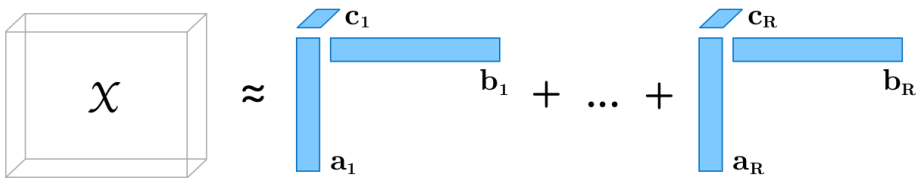
A tenzorszámítással való ismerkedéshez Kolda és Bader (2009) és Rabanser és mtsai (2017) a fő kiindulópontok. Ahogy ezekből is kiderül, nemcsak egyféleképpen lehet általánosítani az SVD alapgondolatát. A következő két szakasz a két legnépszerűbb kiterjesztést, a kanonikus poliadikus felbontást és az általánosabb Tucker-felbontást ismerteti. E két algoritmuscsalád interpretálásának lehetőségeit a jelfeldolgozás és a gépi tanulás kettős szempontjából Sidiropoulos és mtsai (2017) mutatja be.

2.1. Kanonikus poliadikus felbontás

A kanonikus poliadikus felbontás (Canonical Polyadic Decomposition, CPD, más néven CanDecomp, `Para`llel `Fac`tor modell, rangfelbontás vagy Kruskal-felbontás, Carroll és Chang (1970)) a eredeti tenzort 1-rangú tenzorok lineáris kombinációjaként közelíti. Egy 1-rangú tenzor nem más, mint vektorok tenzorszorzata, ugyanúgy ahogy két vektor diádszorzata egy 1-rangú mátrix, lásd az 1-es ábrát.

A váltakozó legkisebb négyzetek algoritmusa (Alternating Least Squares, ALS, Carroll és Chang (1970); Harshman (1970)) iteratív módszer a CPD kiszámítására. Egy-egy iterációban egy híján az összes módot rögzítjük, és a fennmaradót illesztjük. Az ALS nem garantálja a konvergenciát, és még ha az meg is történik, nem észlelhető

[2] Ahogy egy korábbi változat névtelen bírálója megjegyezte, érdekes lehet számos olyan igeosztály szóbeágyazáson alapuló vizsgálata, mint „a thetikus mondatok, egzisztenciális mondatok, aspektusok, határozatlan alanyok. Vajon például megfeleltethetők-e a kapott osztályok valamilyen módon az igei aspektusoknak (pl. igekötős igék a magyarban)? . . . Lehet-e itt szerepe a határozatlan alanyoknak?”



1. ábra: Kanonikus poliadikus felbontás, ábra Rabanser és mtsai (2017)-től

egykönnyen. Felhívjuk viszont a figyelmet az ALS-nak egy viszonylag új továbbfejlesztésére, az Orth-ALS-ra (Sharan és Valiant, 2017), lásd a 3. szakaszt.

2.2. Tucker-felbontás

Noha CPD elterjedtebb a nyelvészetben, röviden bemutatjuk az általánosabb Tucker-felbontást is. A Tucker-felbontás (más néven magasabb rendű SVD, Tucker (1966)) egy kisebb méretű $\mathcal{G}$ magtenzort ad, amit tengelyenként egy-egy mátrixszal megszorozva az eredeti tenzor közelítését kapjuk, lásd a 2-es ábrát. Ha az eredeti tenzor tengelyei

$$\text{alany} \times \text{ige} \times \text{tárgy},$$

akkor a három mátrix sorai az alanyokat, az igéket illetve a tárgyakat beágyazó vektorok, a $\mathcal{G}$ tenzor elemei pedig az előbbi három közötti kölcsönhatások szintjét határozzák meg.

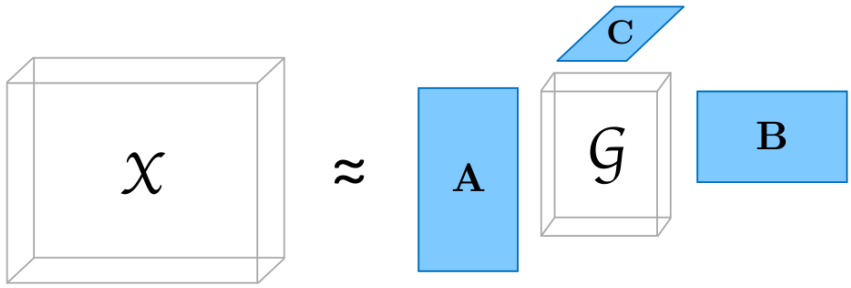
A Tucker-felbontás nem unikus, hiszen $\mathcal{G}$-t az illesztés romlása nélkül transzformálhatjuk, ha a tényezőmátrixokra ugyanannak a transzformációnak az inverzét alkalmazzuk. Az egyediség további követelmények bevezetésével javítható (Kolda és Bader, 2009; Lahat és mtsai, 2015), mint például ritkaság, kis elemek, $\mathcal{G}$ teljes ortogonalitása (*all-orthogonal*), nem-negativitás vagy függetlenség.

2.3. Az együtt-előfordulások számának súlyozása

	korpusz	tengelyek	asszociációs mérték	rang
Van de Cruys (2009)	holland, .5 B	10 K alany × 1 K ige × 10 K direkt tárgy	PPMI	50 . . . 300
Van de Cruys (2011)	holland, .5 B	10 K alany × 1 K ige × 10 K direkt tárgy	kétféle PMI	(nincs felbontás)
Van de Cruys és mtsai (2013)	UKWaC, 2 B	10 K alany × 1 K ige × 10 K tárgy	PMI	300
Jenatton és mtsai (2012)	2 M Wp-cikk	30 K alany × 5 K ige × 30 K direkt tárgy	$\mathbb{P} = 1/(1 + \exp(-\mathbf{s}_i \cdot \mathbf{R}_j \otimes \mathbf{o}_k))$	25, 50, 100
Sharan és Valiant (2017)	Wikipedia, 1.5 B	10 K szó × 10 K szó × 10 K szó	$\log(f + 1)$, $w_i = s_i \oplus v_i \oplus o_i$ normalizálva	100
Bailey és mtsai (2018)	.3 B a Wp-ből	1000-es gyakorisági cut-off	(±eltolt) PPMI, w_i normalizálva	300

1. táblázat. Tenzoros nyelvészeti munkák. A korpuszok méretét többnyire a szavak számában mérve tüntettük fel. A képletekhez némi magyarázatot adunk a szövegben.

A nyelvi gyakoriságok ritka tömböt alkotnak, hiszen a legtöbb szó a legtöbb szóval nem fordul elő együtt empirikusan, és a gyakoriságok sok nagyságrendet ölelnek fel



2. ábra: Tucker-felbontás, ábra Rabanser és mtsai (2017)-től.

(Zipf-törvény avagy hatványeloszlás, Manin (2008); Gittens és mtsai (2017)), ezért ritka tenzorokat érdemes használni a puszta gyakoriságoknál kifinomultabb társítási mértékekkel *(association measure)* benépesítve (*populate)*. Ezekre a mértékekre térünk most rá. Az itt csak hivatkozott nyelvi témájú, tenzorfelbontást alkalmazó munkákat az 1-es táblázat összegzi és a 3. szakasz mutatja be.

A legegyszerűbb választás $\log(f + 1)$, ahol f az együttes előfordulási gyakoriság (Sharan és Valiant, 2017). Jenatton és mtsai (2012) a sokrelációs tanulás kontextusába helyezi az $\langle \text{alany}, \text{ige}, \text{tárgy} \rangle$ hármasok modellezését, és a log-bilineáris modell (Mnih és Hinton, 2007; Mikolov és mtsai, 2013a) súlyozási függvényét alkalmazza. van de Cruys három cikke és Bailey és mtsai (2018) egy információelméleti mérték, a *(pozitív) pontonkénti kölcsönös információ ((positive) pointwise mutual information, (P)PMI)*, háromváltozós általánosítását használja (lásd a 2.4 szakaszt). A pozitivitás azt jelenti, hogy annak érdekében, hogy nagyobb pontszámokat tulajdonítsunk a tényleges együtt-előfordulásoknak, mint a nem-látottaknak, a PMI-nél és a következő bekezdésben bemutatott lexikográfiai mértékeknél is kinullázzuk a negatív elemeket.

Egyes lexikográfiai társítási mértékek is hasznosak lehetnek tenzorfelbontásban. A PMI kétféle háromváltozós általánosításáról a következő szakaszban szólunk. Ezt megelőlegezve bármelyik általánosítást használva magától értődően általánosíthatjuk három változóra a Sketch Engine lexikográfiai szoftverben használt *szembetűnőséget (salience,* Kilgarriff és mtsai (2004)*)* is:

$$\log(f(x, y, z)) \cdot PMI(x, y, z).$$

Kísérletezhetünk a Log-Dice (Rychlý, 2008) általánosításával is:

$$\log \frac{3f(x, y, z)}{f(x) + f(y) + f(z)} + c,$$

ahol c-t úgy választjuk, hogy a Log-Dice értékek nem-negatívak legyenek.

2.4. Többváltozós PMI

Mi más lenne a pontonkénti kölcsönös információ (PMI) többváltozós általánosítása, mint

$$\log \frac{p(x,y,z)}{p(x)p(y)p(z)}, \tag{1}$$

– gondolnánk, de valójában ez csak egy a lehetséges általánosítások közül. Van de Cruys (2011) két pontonkénti asszociációs mértéket is bevezet, amelyeknek a várható értéke a kölcsönös információ (Shannon és Weaver, 1949) egy-egy különböző többváltozós általánosítása: az interakciós információ (McGill, 1954) illetve a teljes korreláció (Watanabe, 1960).

Az interakciós információ a feltételes kölcsönös információ fogalmán alapul:[3]

$$\log \frac{p(x,y)p(x,z),p(y,z)}{p(x,y,z)p(x)p(y)p(z)}$$

A teljes korreláció a változókban levő közös információ mennyiségét számszerűsíti. A pontonkénti változat képlete az 1-es egyenletben látható. Az irodalmat követve (Villada Moirón, 2005; Van de Cruys, 2009; Van de Cruys és mtsai, 2013; Bailey és mtsai, 2018) ebben a cikkben többváltozós PMI alatt (többváltozós pontonkénti) teljes korrelációt értünk.

Van de Cruys (2011) arról számol be, hogy holland kísérleteikben mindkét módszer ki tudott emelni száliens alany–ige–tárgy hármasokat: prototipikus SVO-kombinációkat, például *szavazás képvisel véleményt* és rögzített kifejezéseket. A *játszik* megfelelőjére szűkítve a vizsgálódást azt találják, hogy az interakciós információ prototipikus SVO kombinációkat talál, pl. *zenekar játszik szimfóniát*, míg az elterjedtebb változat, melyet ők specifikus korrelációnak neveznek, a *szerepet játszik* konstrukciót és ennek száliens alanyait taglalja.

2.5. Ritka tenzorok Python3-ban

Az adattudományban és a nyelvtechnológiában a legnépszerűbb szabad szoftverek jelenleg Python 3-on alapulnak, ezért most az itt elérhető tenzorfelbontó csomagokra térünk rá, különös tekintettel a ritka tenzorokra. A fő Python 3 könyvtárak multilineáris algebrához és tenzorfaktorizációkhoz a scikit-tensor-py3 (Nickel és Rol) és a tensorly (Kossaifi és mtsai, 2016). Mindkét könyvtár támogatja bizonyos mértékig sűrű és ritka tenzorok CPD és Tucker-felbontását.

3. A nyelvi többértelműség tenzoros modelljei

A 1 táblázat összefoglalja a nyelvészeti munkák néhány jellemzőjét. Tudomásunk szerint Van de Cruys (2009) vezeti be a nem-negatív tenzorfaktorizációs modellt szelekciós-preferencia-indukcióhoz. Van de Cruys és mtsai (2013) a Kullback–Leibler-divergencia

[3] A pontonkénti változat képlete a szitaformulára emlékeztet, csak fel kell cserélni a számlálót és a nevezőt, hogy matematikai értelemben is mértéket kapjunk.

minimalizálására módosítja a tenzorfaktorizációs modellt, ami szerintük az jobban illeszkedik a hosszú farkú eloszláshoz, amilyeneket a nyelvben is találunk. Ebben a cikkben a főnevek rejtett modelljeit (vagyis a szóvektorokat) előre rögzítik akkori hagyományos együttelőfordulás-alapú módszerrel, ami sajnos korlátozza a tenzorok által amúgy felderíthető harmadrendű struktúra kihasználását. Módszerük lényegi részét, a harmadrendű alany-ige-tárgy interakciók indukcióját, pedig a Tucker-felbontás ihlette.

Jenatton és mtsai (2012) a *sokrelációs tanulás (multi-relational learning)* kontextusában tanulnak szemantikus igereprezentációkat, amely paradigma eredetileg olyan entitásokkal (itt főnevek) foglakozik, amelyek között többféle kapcsolattal (itt ige) állhat fenn, például közösségi hálók, ajánló rendszerek, szemantikus web vagy bioinformatikai adatok. Ebben a paradigmában a kapcsolatok készletét modellezik: a kapcsolatok maguk is hasonlóak lehetnek egymáshoz különféle szempontokból. Kísérleteik során az entitásoknak egyetlen ábrázolása van az összes relációra vonatkozóan. A nyelvi tenzort ⟨alany, ige, közvetlen tárgy⟩ együtt-előfordulásokkal népesítik be.

Polajnar és mtsai (2014) a zaj-kontrasztív becslés módszerét (amit ők *plausibility training*nek hívnak) alkalmazzák tranzitív ige-tenzorok felbontásához. Zhang és mtsai (2014) azt vizsgálják, hogy miként lehet a kézzel létrehozott szemantikai erőforrásokat neurális szóbeágyazásokkal kombinálni az antonimáknak a szinonimáktól való elválasztására, ami közismerten nehéz az eloszlásalapú eszközök számára. A tezaurusz adatait és az eloszlási hasonlóságokat tenzoruk egy-egy szeleteként (táblájaként) fecskendezik be.

A *függvényes (functional)* megközelítésben a szavaknak különböző rendű tenzorok felelnek meg. Egy szóhoz tartozó tenzor rendje összhangban van a szónak egy kategoriális nyelvtanban való típusával. Például a főnevek atomi típusok, amelyeket egy vektor képvisel, és a melléknevek olyan mátrixok, amelyek függvényként működnek (Baroni és Lenci, 2010). A tranzitív ige harmadrendű tenzor. Ebben a megközelítésben probléma, hogy meglehetősen sok paraméter lehet már alacsony dimenziónál is. (Fried és mtsai, 2015) úgy orvosolja ezt a problémát, hogy a tenzorok alacsony rangú közelítését használják.

Hashimoto és Tsuruoka (2015) is mátrixként ábrázolják a tranzitív igéket. Modelljük implicite faktorizál egy tenzort abban az értelemben, ahogy a skip-gram is implicit mátrixfelbontás (Levy és Goldberg, 2014b). A tárgyas igék több jelentését megragadják, és egyértelműsítik őket az argumentumaik alapján. A szabad bővítmények hozzájárulását is vizsgálják.

Cotterell és mtsai (2017) a skip-gram modellt általánosítják tenzorfelbontásként, ami lehetővé teszi beágyazások tanítását gazdagabb, magasabb rendű együtt-előfordulásokból, pl. olyan hármasokból, amelyek a kontextusszónak a fókuszszóhoz képesti helyzetére vonatkozó információt is tartalmaznak, vagy morfológiai információt a kapcsolódó szavak közötti paramétermegosztás érdekében. Negyven nyelven kísérleteznek. Ferraro és mtsai (2017) ezt a modellt használva keretszemantikán alapuló tenzorokat tanítanak. A szemantikus proto-szerepeket (*semantic proto-role*, SPR, Dowty (1991)) egyfajta folytonos keretszemantikának (Fillmore és mtsai, 1976) tekintik, ami bizonyos tulajdonságok valószínűségét ragadja meg, a szerepeket pedig e tulajdonságok csoportjaiként jellemzik. Ferraroék ilyen SPR-alapú várható tulajdonságokat rögzítenek szóbeágyazásokban.

Sharan és Valiant (2017)[4] egy általános szóbeágyazást készít szimmetrikus 3-módú tenzorokból. Cikkük lényege az *Ortogonális ALS (Orth-ALS)*, a ALS megközelítésnek egy olyan módosítása, ami ugyanolyan hatékony, mint a szokásos ALS, de bizonyíthatóan megtalálja a valódi tényezőket véletlen inicializálással a szokásos inkoherenciafeltételezések mellett, azaz hogy a valódi tényezők kevéssé korrelálnak egymással, ami teljesül az NLP-s alkalmazásokban, ahol a közelítő tenzor rangja általában szignifikánsan szublineáris a tér dimenziójában. Az Orth-ALS időről időre „ortogonalizálja" a tényezők becslését, megakadályozva, hogy több kiszámított tényező ugyanazt a valódi tényezőt „üldözze". A szóbeágyazásokat úgy hozzák létre, hogy a három kiszámolt faktormátrixot (egyenként 100 látens dimenzió) konkatenálják egy 300-oszlopos mátrixszá, majd normalizálják a sorokat.

Megemlítjük Bailey és mtsai (2018)[5]-at is, akik egy 3-módú szimmetrikus tenzort képeznek, amely azzal a szójelentés-klaszterezési szempontból figyelemre méltó tulajdonsággal bír, hogy egy alkalmas kontextusvektorral való pontonkénti szorzás segítségével jelentésvektorokat kapnak. Végül Frandsen és Ge (2019) *Szintaktikus RANDWALK* modellje különféle mondattani kapcsolatokat ragad meg egy szóhármasok közötti PMI-t alkalmazó tenzorral.

3.1. Minőségi elemzés a munkákban

⟨athlete, run, race⟩	finish (.29), attend (.27), win (.25)
⟨user, run, command⟩	execute (.42), modify (.40), invoke (.39)
⟨man, damage, car⟩	crash (.43), drive (.35), ride (.35)
⟨car, damage, man⟩	scare (.26), kill (.23), hurt (.23)

2. táblázat. Tranzitív szerkezetben kontextualizált igékhez leghasonlóbb igék Van de Cruys és mtsai (2013)-nál.

A Van de Cruys (2009) által végzett kvalitatív kiértékelés a látens dimenziók elemzésén alapul: az egyes dimenziókat az azokban legnagyobb abszolút értékű koordinátát kapó alanyok, igék és tárgyak szerint értelmezik. Úgy találják, hogy a 100 dimenzió közül 44 keretszemantikát példáz. Egy olyan dimenzióban, amit úgy hívhatunk, hogy *rendőrség letartóztat gyanúsítottat*, a legnagyobb súlyú alanyok, igék és tárgyak olyan szavak, mint például *rendőrség*, *letartóztat* illetve *gyanúsított*. További példák: *többség támogat javaslatot* vagy *kormány küld csapatot.* További 43 dimenzió szemantikája kevésbé egyértelmű: ezek egyetlen igét képviselnek, esetleg egy ige különféle jelentései keverednek. Tizenhárom rejtett dimenzió konkrét igei szerkezeteket tartalmaz, például x *játszik szerepet*, ahol az alanyi oszlop egyenletesen oszlik meg több tucat szó között, pl. *bosszú, szégyen, intézmény*, *kultúra* vagy *osztódás*.

[4] http://web.stanford.edu/~vsharan/orth-als.html

[5] https://github.com/popcorncolonel/tensor_decomp_embedding

Van de Cruys és mtsai (2013) tenzorában a szeletek igéket képviselnek. Ők úgy szemléltetik az adatokat, hogy hármasokhoz a bennük szereplő, kontextualizált igékhez leghasonlóbb igéket mutatják meg, lásd a 2-es táblázatot.

A Frandsen és Ge (2019) kvalitatív kiértékelésében együttes melléknév-főnév illetve ige-tárgy vektorokhoz legközelebbi beágyazású szavakat néznek.

3.2. Számszerű elemzés a munkákban

Van de Cruys (2009) ál-egyértelműsítési feladatban értékeli ki a modelljét, ahol azt kell megítélni, hogy melyik alany (s vagy s') és közvetlen tárgy (o vagy o') valószínűbb egy adott v ige esetében. A tesztkészletet úgy építi fel, hogy $\langle s, v, o\rangle$ -t a korpuszból veszi, míg s' és o' egy-egy véletlenszerűen választott alany illetve közvetlen tárgy a korpuszból, például *fiatal/koalíció iszik sört/részvényt*. Tudomásunk szerint a tesztkészletük nem érhető el.

Grefenstette és Sadrzadeh (2011) tárgyas igék egyértelműsítésére vonatkozó adathalmaza igepárokat tartalmaz egy-egy alannyal és tárggyal. A feladat az igék többértelműségén alapszik (Kartsaklis és Sadrzadeh, 2013; Milajevs és mtsai, 2014; Polajnar és mtsai, 2014). Például az angol *meet* ige többértelmű; egyik jelentésében a *satisfy*-hoz hasonlít, egy másikban a *visit*-hez: A *beach meet standard* kontextusban a *satisfy*-hoz és csak ahhoz, a *representative meet official* környezetben pedig a *visit*-hez. A feladat ezen hasonlóságok predikciója. Van de Cruys és mtsai (2013) ezeken a tárgyas mondatokon értékelik ki számszerűen a rendszerüket.

Kartsaklis és Sadrzadeh (2013) egy másik teszthalmazt, Mitchell és Lapata (2010) $\langle$ige, tárgy$\rangle$ szerkezetek hasonlóságára vonatkozó adatát egészíti ki: az eredeti párokat alanyokkal látja el úgy, hogy a hasonlóság mértékét igyekeznek őrizni, hogy az emberi hasonlóságítéletek érvényesek maradjanak. Kartsaklis és Sadrzadeh (2014) olyan változatát adja közre az adathalmaznak,[6] ahol már a hármasokat értékeltetik ki Amazon Türkkel. Polajnar és mtsai (2014), Fried és mtsai (2015) és Hashimoto és Tsuruoka (2015) Grefenstette-ék adathalmazán és ez(ek)en értékelnek ki.

Jenatton és mtsai (2012) két feladatban értékelik ki modelljeiket: adott alanyhoz és közvetlen tárgyhoz jósolják be a megfelelő igét, illetve lexikai hasonlósági osztályozást végeznek. Ők is közzéteszik a tesztadatot, de mi azt találtuk, hogy hármasaik kissé zajosabbak pl. Grefenstette-ékéinél.

Noha nem triviális, hogy az antonímia jobban támaszkodik-e a hárommódú együttes előfordulásokra, mint például a szinonímia, Zhang és mtsai (2014) GRE antonim kérdésekben (Mohammad és mtsai, 2008) értékelik ki a munkájukat. A teszthalmazról Zhangék azt írják, hogy az adatkészlet „szemmel láthatóan köztulajdonban" van, és kérésre elérhető.

A szóvektorok tesztelésének egyik legnépszerűbb módszere a szóhasonlósági rangsorolási feladat (Cotterell és mtsai (2017) is így értékelnek ki), kiváltképp a SimLex-999 (Hill és mtsai, 2014). Noha a szópárok hasonlósága nem közvetlenül a háromirányú interakciókat célozza meg, úgy gondoljuk, hogy a SimLex-999 igei megfelelője, a SimVerb (Gerz és mtsai, 2016), sőt az alanyok és a tárgyak tekintetében maga a SimLex-999

[6] http://www.cs.ox.ac.uk/activities/compdistmeaning/

is, hasznos józanság-ellenőrzést *(sanity check)* nyújt a tenzorfelbontó modellek számára is.

Sharan és Valiant (2017) nem teszteli kifejezetten a >2-rendű kapcsolatok modellezését, hanem szokásos szóanalógiában („a *kutyá*hoz úgy viszonyul a *kan*, mint a *macská*hoz a(z) x") és szemantikai szóhasonlósági feladatokban értékelik ki az ortogonalizált tenzorként nyert beágyazásokat. Azzal, hogy az Orth-ALS-t használják a szokásos ALS helyett, jelentős javulást kapnak, ám a mátrix-SVD módszer továbbra is felülmúlja a tenzoralapú módszereket. Miután felvetik azt a pesszimista magyarázatot, miszerint a természetes nyelv esetleg nem tartalmaz eléggé gazdag magasabb rendű függőségeket a szűk kontextusban megjelenő szavak között a 2-módú szerkezeten túl, másodikként azt a lehetséges magyarázatot adják a gyenge teljesítményre, hogy csak a két vizsgált feladathoz nem szükséges ez a fajta magasabb rendű statisztika. Végül Frandsen és Ge (2019) a már említett melléknév–főnév szókapcsolatok hasonlóságára vonatkozó feladatban (Mitchell és Lapata, 2010) értékelik ki a munkájukat.

3.3. Magyar munkák

Bár a cikkünk angol tárgynyelvű, a magyar konferencia közönsége számára érdekes lehet a kapcsolódó magyar munkák bemutatása – a teljesség leghalványabb igénye nélkül. A magyar nyelvű szójelentés-osztályozás *(word sense disambiguation)* a gépi tanulás szempontjából legalább Miháltz (2005)-ig és Vincze és mtsai (2008)-ig nyúlik vissza. Az igéknek sok kutató szentelte a figyelmét nyelvészektől a szűkebb értelemben vett nyelvtechnológusokig (Dressler és Ladányi, 2000; Kuti és mtsai, 2010; Miháltz és Sass, 2013).

A magyar igei konstrukciók fő adatbázisai a megjelenés sorrendjében a Mazsola (Sass, 2015, 2018), a Tádé (Kornai és mtsai, 2016) és a Manócska (Kalivoda és mtsai, 2018; Kalivoda, 2019). A magyar szóbeágyazós munkák közül Makrai (2015); Siklósi (2016); Berend (2018); Kardos és mtsai (2019)és Döbrössy és mtsai (2019) munkáit emeljük ki, mint cseppeket a tengerből.

4. Következtetés és a jövőbeni kutatás

Összességében elmondhatjuk, hogy a tenzorfelbontás a fősodorra (Hewitt és Manning, 2019) merőleges irányt kínál a nyelvi szerkezet adatvezérelt megértésében. Hosszú távon, amint azt az 1. szakaszban már említettük, szemantikai igeosztályokat szeretnénk felügyeletlenül tanulni. Ha az igei beágyazás-vektorok Levin (1993)-féle igeosztályoknak megfelelő klaszterekbe rendeződnek, akkor a többértelmű igéket a klaszterekből kimaradó vektorok formájában azonosíthatjuk be. Ez a kutatási vonal a többnyelvű paradigmába is kiterjeszthető (Vulić és mtsai, 2017; Majewska és mtsai, 2018; Sun és mtsai, 2010).

Köszönetnyilvánítás

Hálás vagyok Tülay Adalınak, aki a 2018-as DeepLearn nyári egyetem lelkesítő előadójaként felhívta a figyelmemet a tenzorfelbontás általi interpretációban rejlő lehetőségekre, valamint Berend Gábornak, Borbély Gábornak, Indig Balázsnak, Kalivoda

Ágnesnek, Kornai Andrásnak, Sass Bálintnak, Simon Eszternek, Szécsényi Tibornak és korábbi változatok névtelen bírálóinak (MSZNY 2019, ACL 2019, Repl4NLP 2019, Maleczki 65) hasznos megjegyzéseikért. Kutatásomat részben a 2018-1.2.1-NKP-2018-00008 *A mesterséges intelligencia matematikai alapjai* és az NKFIH 120145-ös *Szószerkezet felismerése mélytanulással* projekt támogatta.

Hivatkozások

Bailey, E., Meyer, C., Aeron, S.: Learning semantic word respresentations via tensor factorization (2018), https://openreview.net/forum?id=B1kIr-WRb, arXiv:1705.08968 [cs.AI]

Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics. Computational Linguistics 36(4), 673–721 (2010)

Berend, G.: Towards cross-lingual utilization of sparse word representations. In: Vincze, V. (szerk.) XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). pp. 272–280. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2018)

Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika 35, 283–319 (1970)

Cotterell, R., Poliak, A., Durme, B.V., Eisner, J.: Explaining and generalizing skip-gram through exponential family pca. In: ACL (2017)

Van de Cruys, T.: A non-negative tensor factorization model for selectional preference induction. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. pp. 83–90. Association for Computational Linguistics, Athens, Greece (Mar 2009), https://www.aclweb.org/anthology/W09-0211

Van de Cruys, T.: Two multivariate generalizations of pointwise mutual information. In: Proceedings of the Workshop on Distributional Semantics and Compositionality. pp. 16–20. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://www.aclweb.org/anthology/W11-1303

Van de Cruys, T., Poibeau, T., Korhonen, A.: A tensor-based factorization model of semantic compositionality. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1142–1151. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://www.aclweb.org/anthology/N13-1134

Döbrössy, B., Makrai, M., Tarján, B., Szaszák, G.: Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). pp. 187–193. Association for Computational Linguistics, Florence, Italy (Aug 2019), https://www.aclweb.org/anthology/W19-4321

Dowty, D.: Thematic proto-roles and argument selection. Language 67(3), 547–619 (1991)

Dressler, W.U., Ladányi, M.: Productivity in word formation (wf): a morphological approach. Acta Linguistica Hungarica 47(1-4), 103–145 (2000)

Ferraro, F., Poliak, A., Cotterell, R., Durme, B.V.: Frame-based continuous lexical semantics through exponential family tensor factorization and semantic proto-roles. In: Joint Conference on Lexical and Computational Semantics (*SEM) (2017)

Fillmore, C.J., és mtsai: Frame semantics and the nature of language. In: Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech. pp. 20–32 (1976)

Frandsen, A., Ge, R.: Understanding composition of word embeddings via tensor decomposition. In: 7th International Conference on Learning Representations, ICLR 2019 (2019), https://openreview.net/forum?id=H1eqjiCctX, arXiv preprint arXiv:1902.00613

Fried, D., Polajnar, T., Clark, S.: Low-rank tensors for verbs in compositional distributional semantics. In: ACL (2015)

Gerz, D., Vulić, I., Hill, F., Reichart, R., Korhonen, A.: SimVerb-3500: A large-scale evaluation set of verb similarity. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2173–2182. Association for Computational Linguistics, Austin, Texas (Nov 2016), https://www.aclweb.org/anthology/D16-1235

Gittens, A., Achlioptas, D., Mahoney, M.W.: Skip-gram – zipf + uniform = vector additivity. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 69–76. Association for Computational Linguistics (2017), http://aclweb.org/anthology/P17-1007

Grefenstette, E., Sadrzadeh, M.: Experimenting with transitive verbs in a DisCoCat. In: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics. pp. 62–66. Association for Computational Linguistics, Edinburgh, UK (Jul 2011), https://www.aclweb.org/anthology/W11-2507

Harshman, R.A.: Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics 16, 1–84 (1970), http://publish.uwo.ca/~harshman/wpppfac0.pdf

Hashimoto, K., Tsuruoka, Y.: Learning embeddings for transitive verb disambiguation by implicit tensor factorization. In: 3rd Workshop on Continuous Vector Space Models and their Compositionality (2015)

Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4129–4138 (2019)

Hill, F., Reichart, R., Korhonen, A.: Multi-modal models for concrete and abstract concept meaning. Transactions of the Association for Computational Linguistics 2(10), 285–296 (2014)

Jenatton, R., Roux, N.L., Bordes, A., Obozinski, G.: A latent factor model for highly multi-relational data. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2. pp. 3167–3175. NIPS'12, Curran Associates Inc., USA (2012), http://dl.acm.org/citation.cfm?id=2999325.2999488

Kalivoda, A.: Véges erőforrás végtelen sok igekötős igére [A finite resource for infinitely many Hungarian particle verbs]. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 331–344 (January 2019)

Kalivoda, Á., Vadász, N., Indig, B.: MANÓCSKA: A Unified Verb Frame Database for Hungarian. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (szerk.) Proceedings of

the 21st International Conference on Text, Speech and Dialogue—TSD 2018, Brno, Czech Republic. Lecture Notes in Artificial Intelligence, vol. 11107, pp. 135–143. Springer-Verlag (Sep 2018)

Kardos, P., Berend, G., Farkas, R.: Kísérletek tudásbázis- és mondatkörnyezet-alapú beágyazásokkal magyar nyelvre. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 153–162. Szegedi Tudományegyetem, Informatikai Intézet (2019)

Kartsaklis, D., Sadrzadeh, M.: Prior disambiguation of word tensors for constructing sentence vectors. In: EMNLP (2013)

Kartsaklis, D., Sadrzadeh, M.: A study of entanglement in a categorical framework of natural language. In: The 11th workshop on Quantum Physics and Logic (6 2014), arXiv:1412.8102

Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: Sketch engine. In: Williams, G., Vessier, S. (szerk.) Proceedings of Euralex. pp. 105–116. Lorient, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines (July 2004)

Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM review 51(3), 455–500 (2009)

Kornai, A., Nemeskey, D.M., Recski, G.: Detecting optional arguments of verbs. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 2815–2818. European Language Resources Association (ELRA), Paris, France (may 2016)

Kossaifi, J., Panagakis, Y., Anandkumar, A., Pantic, M.: Tensorly: Tensor learning in python. Journal of Machine Learning Research (JMLR) 20, 1–6 (2016), arXiv preprint arXiv:1610.09555

Kuti, J., Héja, E., Sass, B.: Sense disambiguation – „ambiguous sensation"? evaluating sense inventories for verbal wsd in hungarian. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (szerk.) Proceedings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-)Eastern European Languages. pp. 23–30. European Language Resources Association (ELRA) (2010)

Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE 103(9), 1449–1477 (2015)

Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104(2), 211 (1997)

Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press (1993)

Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 302–308. Association for Computational Linguistics, Baltimore, Maryland (June 2014a), http://www.aclweb.org/anthology/P14-2050

Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (szerk.) Advances in Neural Information Processing Systems 27. pp. 2177–2185 (2014b)

Majewska, O., Vulić, I., McCarthy, D., Huang, Y., Murakami, A., Laippala, V., Korhonen, A.: Investigating the cross-lingual translatability of VerbNet-style classification. Language Resources and Evaluation 52(3), 771–799 (2018)

Makrai, M.: Comparison of distributed language models on medium-resourced languages. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). pp. 22–33. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2015)

Manin, D.Y.: Zipf's law and avoidance of excessive synonymy. Cognitive Science 32, 1075–1098 (2008)

McGill, W.: Multivariate information transmission. Transactions of the IRE Professional Group on Information Theory 4(4), 93–111 (1954)

Miháltz, M.: Towards a hybrid approach to word-sense disambiguation in machine translation. In: Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., Nikolov, N. (szerk.) RANLP-2005 Workshop: Modern Approaches in Translation Technologies (September 2005)

Miháltz, M., Sass, B.: What do we drink? automatically extending hungarian wordnet with selectional preference relations. In: Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora. pp. 105–109 (2013)

Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (szerk.) 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings (May 2013a), http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (szerk.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013b)

Milajevs, D., Kartsaklis, D., Sadrzadeh, M., Purver, M.: Evaluating neural word representations in tensor-based compositional settings. In: EMNLP. pp. 708–719 (2014)

Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive Science 34, 1388–1429 (2010)

Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: Proceedings of the 24th international conference on Machine learning. pp. 641–648. ACM (2007)

Mohammad, S., Dorr, B., Hirst, G.: Computing word-pair antonymy. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 982–991. Association for Computational Linguistics (2008)

Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics (2014), http://www.aclweb.org/anthology/D14-1162

Polajnar, T., Rimell, L., Clark, S.: Using sentence plausibility to learn the semantics of transitive verbs. In: NIPS Learning Semantics Workshop (2014), in arXiv, some minor errata fixed.

Rabanser, S., Shchur, O., Günnemann, S.: Introduction to tensor decompositions and their applications in machine learning (2017), http://arxiv.org/abs/1711.10781v1, arXiv:1711.10781 [stat.ML]

Rychlý, P.: A lexicographer-friendly association score. In: Proceedings of Recent Advances in Slavonic Natural Language Processing. pp. 6–9 (2008)

Sass, B.: A lattice based algebraic model for verb centered constructions. In: TSD. pp. 231–238. Springer (2018)

Sass, B.: 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet [28 million syntactically analyzed sentences and 500 000 verb constructions in Hungarian]. In: Attila, T., Viktor, V., Veronika, V. (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). pp. 303–308. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2015)

Shannon, C.E., Weaver, W.W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)

Sharan, V., Valiant, G.: Orthogonalized ALS: A theoretically principled tensor decomposition algorithm for practical use. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. pp. 3095–3104 (August 2017), http://proceedings.mlr.press/v70/sharan17a.html

Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C.: Tensor decomposition for signal processing and machine learning. IEEE Transactions on signal processing 65(13), 3551–3582 (Jul 2017), https://doi.org/10.1109/TSP.2017.2690524

Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 115–126. Springer (2016)

Sun, L., Korhonen, A., Poibeau, T., Messiant, C.: Investigating the cross-linguistic potential of VerbNet: style classification. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1056–1064. Association for Computational Linguistics (2010)

Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika 31(3), 279–311 (1966)

Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 141–188 (2010)

Villada Moirón, M.B.: Data-driven identification of fixed expressions and their modifiability. Ph.D.-értekezés, University of Groningen (2005)

Vincze, V., Szarvas, G., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, C., Csirik, J.: Hungarian word-sense disambiguated corpus. In: Calzolari, N., ChoukriK, Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D. (szerk.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). pp. 3344–3349. European Language Resources Association (ELRA) (2008)

Vulić, I., Mrkšić, N., Korhonen, A.: Cross-lingual induction and transfer of verb classes based on word vector space specialisation. arXiv preprint arXiv:1707.06945 pp. 2546–2558 (Sep 2017), https://www.aclweb.org/anthology/D17-1270

Watanabe, S.: Information theoretical analysis of multivariate correlation. IBM Journal of research and development 4(1), 66–82 (1960)

Zhang, J., Salwen, J., Glass, M., Gliozzo, A.: Word semantic representations using Bayesian probabilistic tensor factorization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1522–

1531. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://www.aclweb.org/anthology/D14-1161

A természetesnyelv-feldolgozás fizikai és nyelvi határai

Mészáros Evelin

Clementine
1115 Budapest Hungary
emeszaros@clementine.hu

Kivonat: A természetesnyelv-feldolgozó rendszerek fejlettsége mára már elért egy olyan szintet, hogy lehetőségünk nyílik különböző szoftverek segítségével olyan rendszereket építeni, amelyek megpróbálják értelmezni a feldolgozandó szöveges információt. A rendelkezésünkre álló eszközök egyre növekvő tárháza, amelyhez a nyílt forráskódú programnyelvek térhódítása is hozzájárul, lehetővé teszi azt a helyzetet, hogy egy másik idegen nyelvre írt módszertant a saját anyanyelvünkön implementáljunk. Ennek azonban lehetnek nehézségei, melyek a feldolgozandó nyelv jellegzetességeitől függenek, erre viszont a legtöbb esetben a szerzők nem adnak útmutatást. Jelen tanulmány ilyen útmutatásokat kíván nyújtani, vagyis hogyan tudunk egy angol nyelvre speciálisan kifejlesztett módszertant átültetni magyar nyelvre?

Kulcsszavak: szemantikai hasonlóság, ontológia, NLP

1 Bevezetés

A természetesnyelv-feldolgozáson belül többféle alkalmazási területen lehet szükség két szöveg összehasonlítására. Számos forrás áll rendelkezésünkre, amelyben egy ilyen feladat megoldására javasolnak módszereket a szerzők: ezek közül megemlíthetjük a kulcsszókinyerést vagy akár a kivonatolást, de ezek a metódusok rövid szövegek összehasonlításánál kudarcot vallanak, hiszen nagyon kicsi az esély arra, hogy az adott rövid leírásban szerepel akár egy kulcsszó is. Ebben az esetben szükségünk van egy olyan módszertanra, amely túllép az egyszerű gyakoriságalapú megközelítésen, és figyelembe veszi az adott nyelv struktúráját és a benne szereplő szavak egymáshoz kapcsolódó viszonyait is.

Ahhoz, hogy a gyakoriságok segítségével jól tudjuk definiálni egy adott szó környezetét, hatalmas méretű témaspecifikus korpuszra van szükségünk az adott nyelven – ennek hiánya, illetve idő- és szaktudásigényes előállítása az oka a hagyományos értelemben vett szóvektorokkal való ábrázolás elvetésének is –, ezért lehet egy jó megközelítés a nyelv hálózatként való definiálása. A hálózati megközelítés a nyelvfeldolgozás és a statisztika területén is megjelent, és számos tanulmány mutat rá arra, hogy a jelentésbeli hasonlóságot egy egész rendszerhez viszonyítva érdemes meghatározni. Ehhez azonban szükség van egy forrásra, amely tartalmazza a szavak egymáshoz viszonyított

kapcsolatát is – erre lehet egy jó kiindulási pont az ún. Wordnet[1], amely számos nyelven elérhető többé-kevésbé kidolgozott formában.

A magyar nyelv helyzete nagyon egyedi – agglutináló nyelv révén a toldalékok azonosítása korántsem olyan triviális, mint egyéb (pl. angol) nyelvek esetén, ahol a toldalékok nem a szavakkal egybeírva, hanem többnyire azok előtt különállóan helyezkednek el. További nehézséget jelenthet a szabad szórend is szemben egyéb nyelvek (pl. angol) kötött szórendjével.

A nyílt forráskódú szoftverek terjedésével együtt az autodidakta nyelvészek és adatelemzők is számos opciót mutatnak arra vonatkozóan, hogy hogyan tudjuk két szövegrészről eldönteni, hogy hasonló jelentésűek-e vagy sem. Ezeket a módszereket széles körben alkalmazzák sokféle tudásbázisalapú alkalmazásnál, egyéb információkinyerő rendszerek esetén, érzelemdetektáláshoz vagy akár a biostatisztikában is (Slimani, 2013). A lehetőségek tárháza azt a látszatot keltheti a kutatóban, hogy ezeket a modelleket és forrásokat ugyanebben a formában lehetséges reprodukálni.

A tanulmányban azt szeretném bemutatni, hogy a számos nagy világnyelvekre (legfőképpen angolra) készült szemantikai hasonlóságalapú módszerek milyen szempontok figyelembevételével ültethetők át egy adott nyelvre. Az írás második fejezetében a szemantikai hasonlóság elméleti hátterét és számításának néhány alternatíváját fogom részletezni, majd a következő fejezetben a felhasznált módszertant fogom leírni. A harmadik fejezetben pedig iránymutatást szeretnék adni arra vonatkozóan, hogy milyen szempontokat érdemes figyelembe venni egy idegen nyelvű módszertan magyar nyelvre való átültetésekor.

2 Elméleti háttér

2.1 Szemantikai hasonlóság

A bevezetőben utaltam már rá, hogy számos területen szükségünk lehet arra, hogy két szövegről eldöntsük, jelentésükben hasonlítanak-e egymásra vagy sem. Például ha egy szöveg koherenciáját szeretnénk megállapítani (Lapata és Barzilay, 2005), vagy ha egy gépi fordítás eredményét szeretnénk automatikusan értékelni (Papineni és mtsai, 2002), de akkor is, ha egy szövegnek szeretnénk a lényeges tartalmát kinyerni automatizáltan (Salton és mtsai, 1997). A mesterséges intelligencia térhódításának korában azonban akár egy párbeszéd összeállításánál is felmerülhet az igény az alkalmazására, vagy kérdés-válasz párok esetében is hasznos lehet ez a tudás.

A szemantikai hasonlóság legegyszerűbb megközelítése szerint a hasonlósági mutatót aszerint számoljuk, hogy mennyi azonos szó szerepel a két összehasonlítandó szövegben. (Mihalcea és mtsai, 2006). Számos továbbfejlesztésre történt kísérlet további szempontok bevonásával mint például a szótövezés, stopszó-eltávolítás, szófaji egyértelműsítés, leghosszabb összeillő tagmondatpár megválasztása, illetve egyéb súlyozó és normalizáló tényezők figyelembe vétele (Salton és mtsai, 1997).

[1] https://wordnet.princeton.edu

A legtöbb esetben ún. word-to-word összehasonlítások esetén a szövegben szereplő szavakat vesszük elemzési egységnek és ezeket hasonlítjuk össze egymással, viszont sokszor szükséges lehet az is, hogy többtagú kifejezéseket is összetartozónak vegyünk.

Jelen tanulmány elméleti hátterét egy olyan írás (Li és mtsai, 2006) adja, amely korpusz statisztikák és szemantikai hálózatok együttes figyelembevételével számolja ki két rövid szöveg közötti hasonlóság mértékét. Egyetértek az említett szerzőkkel abban, miszerint három fő hátránya van a szemantikai hasonlóság kiszámításának, amelyek egyben alkalmazási nehézségnek is felfoghatók (Li és mtsai, 2006):

- egy mondat szavai csak nagyon sokdimenziós vektortérben ábrázolhatók, és ez alacsony modellteljesítménnyel társulhat
- a legtöbb módszertan a kutató intenzív beavatkozását igényli a szövegelőkészítés folyamatában
- egy létrehozott modell nem adaptálható könnyen egyéb témakörökre

A szerző (Li és mtsai, 2006) megfogalmazza azt az igényt, miszerint egy hatékony szövegösszehasonlító módszer csak a mondatok értelmére koncentráljon, képes legyen automatikusan bővülni a kutató kézi beavatkozása nélkül (vagy csak korlátozott kutatói beavatkozással), és könnyen adaptálható legyen egyéb témakörökre is.

Hangsúlyozom, hogy jelen tanulmányban bemutatott módszertan kifejezetten rövid szövegek összehasonlítására mutatott hatékony eredményeket mind angol mind pedig magyar nyelven, hosszú szövegek esetén egyéb módszerek is hatékonynak bizonyulhatnak, melyekre terjedelmi korlátok miatt itt nem térek ki.

A következő alfejezetben a szemantikai hasonlóság kiszámításának azt a módját fogom ismertetni, melyet a későbbiekben leírt módszertanhoz is alkalmazok.

2.2 Szövegből adat - a szöveg reprezentációja

Egy szöveg vagy mondat matematikai modellje egy vektor, mely a tisztított (stopszómentes, lemmatizált) mondatban szereplő szavak előfordulását vagy előfordulási gyakoriságát tartalmazza.

Két mondat az 1. táblázatban látható módon ábrázolható vektorként:

1. táblázat: Mondatok vektorreprezentációja

	pénzintézet	van	hitel	bank	rendelkezik	kölcsön
A Egy pénzintézetnél van hitele. [	1	1	1	0	0	0]
B Egy banknál rendelkezik kölcsönnel. [	0	0	0	1	1	1]

Két szöveg (ebben az esetben mondat) hasonlóságának, távolságának legelterjedtebb mértéke a koszinusz távolság, melyet az 1-es képletben szereplő módon lehet kiszámítani, és amelyben A és B a mondatokat leíró vektorok.

$$Similarity(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad \textbf{(1)}$$

Ha a két mondatban nincsenek közös szavak, akkor a hasonlóság értéke 0, azaz a két mondat ezen mérték szerint egyáltalán nem tekinthető hasonlónak. Az 1. táblázatban szereplő példa is jól illusztrálja, hogy habár a két mondat szókészletben különbözik ugyan, de tartalmilag szinte teljesen megegyezik. A legtöbb publikációban található

módszertan gyengesége meglátásom szerint pedig pont abban rejlik, hogy csak a formailag teljesen megegyező szavakat – elemzési egységeket - tudja azonosítani, így az adott nyelvrendszer hálózatként való értelmezése erre a problémára egy hatékony kezelési eljárás lehet, melyet a 3.1 fejezetben fogok részletesen bemutatni.

3 A módszertan bemutatása

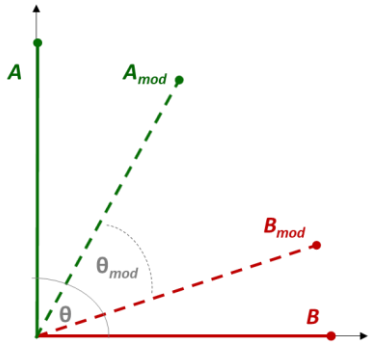
3.1 A felhasznált módszertan alapjai[2]

A módszertan kiválasztásánál fontos volt figyelembe venni azt is, hogy milyen egyéb módszerek lehetségesek az adott problematika megoldására. A kiinduló kérdés az volt, hogyan lehetne hatékonyan eldönteni, hogy két rövid terjedelmű magyar nyelvű mondat hasonlónak tekinthető-e. Tovább bonyolította a helyzetet az is, hogy nem egy általános témakörről lett volna szó, hanem egy konkrét szűkebb témáról, amelynek speciális nyelvezete van. A lehetőségek közül a divatos neurális hálók módszere nem tűnt alkalmazhatónak, hiszen a témának megfelelően annotált magyar nyelvű korpusz tudomásom szerint még nem áll rendelkezésre, továbbá az említett felügyelt tanulási algoritmus működésének ún. „fekete dobozként" fogja fel a tudományos szféra és a működésbe való kutatói beavatkozás túlságosan bonyolult. A terjedelmes dokumentumok esetén alkalmazott kulcsszókinyerés látszólag itt azért sem volt alkalmazható, mert a szövegek rövidsége miatt a szavak gyakorisága alapján nem lehetne kulcsszavakat azonosítani vagy azok lényegében egyenlőek lennének a mondat szókészletével, és ez folyamatos és aktív kutatói frissítését követelné a létrehozott modellnek. Korpusz hiányában a topik-modellezési eljárások sem látszottak hatékonynak.

Az alapul választott módszertan (Li és mtsai, 2006) erősségét az adja, hogy ötvözi a tudásbázis- és a korpuszalapú megközelítést. A módszertanban rendelkezésre álló tudásbázis egy hálózatként definiálható, ahol a hálózat pontjai az elemzési egységek – alapesetben szavak -, és hálózatelméleti alapfogalmakkal lehetséges leírni az egységek közötti kapcsolatokat.

A 2.2 fejezetben leírt vektorreprezentáció módosítása adja a módszer egyik sarkalatos pontját. A szöveghasonlóság módszertanához forrásként felhasznált cikk (Li és mtsai, 2006) javaslata szerint ezért a mondatokat leíró vektorok ne azt tartalmazzák, hogy egy adott szó előfordul-e a mondatban vagy sem (0/1), hanem azt, hogy az adott szóhoz milyen mértékben hasonló szó szerepel a mondatban ([0-1] közötti hasonlósági érték). Így a szókészlet alapján teljesen különböző mondatok vektorai „közelíthetők" egymáshoz, amit az 1. ábra szemléltet.

[2] A 3.1 fejezetben bemutatott módszertan az alábbi irodalom alapján készült: Yuhua és mtsai (2006)



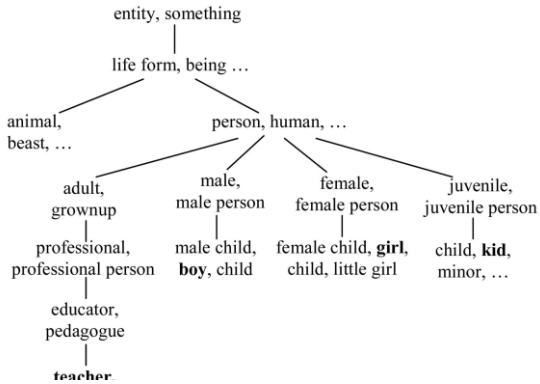
1. ábra: Az A és a B mondat vektorreprezentációjának ábrázolása

A hasonlósági értékekkel módosított vektor szemantikai vektornak nevezhető, amely az 1. táblázat példamondatai esetében a következőképpen néz ki:

2. táblázat: Arányokkal módosított szemantikai vektor

		pénzintézet	van	hitel	bank	rendelkezik	kölcsön	
A	Egy pénzintézetnél van hitele.	[1	1	1	0.9	1	1	]
B	Egy banknál rendelkezik kölcsönnel.	[0.9	0	0	1	1	1	]

Az eredeti vektor értékei módosultak azzal, hogy az értékek helyén nemcsak egyszerű bináris értékek (szerepel-e benne vagy sem), hanem arányszámok szerepelnek, amelyek a szavak egymáshoz viszonyított kapcsolatát tükrözik egy hierarchikus hálózatban – ún. ontológiában. Angol nyelven ugyan rendelkezésre áll egy nyilvánosan használható ontológia, WordNet (wordnet.princeton.edu), magyar nyelven azonban csak általános és jogi témában létezik nyilvánosan elérhető ontológia. További jelentős



2. ábra: Részlet az angol Wordnetből

erőforrásokat igénylő feladat ezért a kutatási kérdésnek megfelelő témaspecifikus hálózat létrehozása, de ezen hálózat létrehozásának folyamata nem képezi jelen tanulmány részét. Két szó hasonlóságát a 2-es képlet szerint írható le:

$$s(w_1, w_2) = f(l, d) = f_1(l) \cdot f_2(d) \quad \textbf{(2)}$$

A 2-es képlet szerint két szó hasonlósága függ attól, hogy mekkora a köztük levő legrövidebb út hossza az ontológiában (l ~ length), illetve attól is, hogy a két szó közös őse milyen mélyen helyezkedik el a hálózatban (d~depth). A 2. ábra szerint (melyen egy részlet látható az angol WordNet ontológiából) az angol *boy* és *girl* szavak között a legrövidebb út hossza 4 egység, míg a közös ősük (*person*) a hálózatban 2 mélységben található, ha az ontológia gyökerének az *entity* pontot vesszük. Li és mtsai (2006) szerint az ontológiából meghatározott úthossz és mélység a 3-as képletben szereplő függvények szerint számolható, ahol α és β a kutató által meghatározott értékek.

$$f_1(d) = \frac{e^{\beta d} - e^{-\beta d}}{e^{\beta d} + e^{-\beta d}} \quad \text{és} \quad f_2(l) = e^{-\alpha l} \quad \textbf{(3)}$$

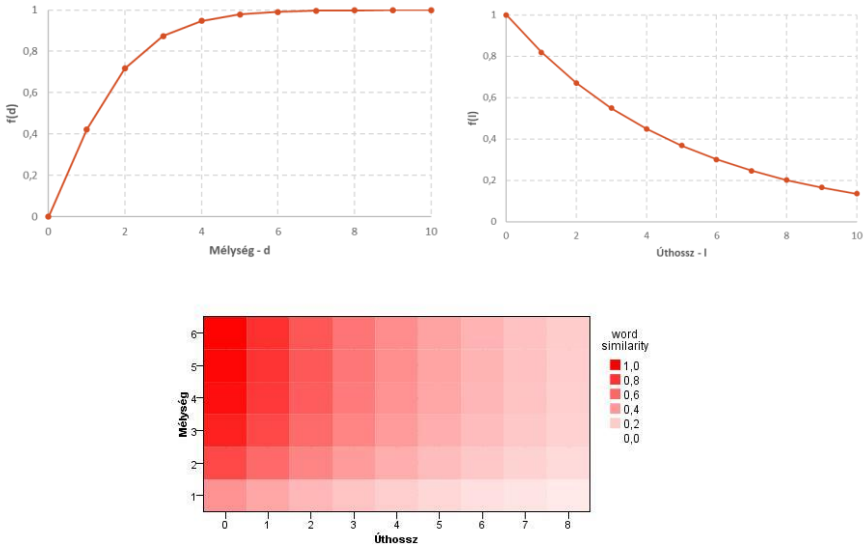
A módszertan magyar nyelvre való átültetése során az α és β paraméterek finomhangolása is megtörtént, de a tesztelés során ezeknek a paraméterek beállítása maradt α=0.3 és β=0.45 (Li és mtsai, 2006).

Két szó annál hasonlóbb egymáshoz, minél rövidebb úton juthatunk el egyik szótól a másikig, valamint minél mélyebben helyezkednek el a hálózatban, hiszen annál pontosabban lehet definiálni a szavak jelentését. A számításnál figyelembe vett úthossz és mélység függését a 3. ábra mutatja.

Két mondat hasonlóságát pedig a 3.1 fejezetben leírt módosítások alapján feltöltött vektorokkal számíthatjuk ki a koszinusz hasonlóság segítségével, melynek programkódrészlete a 4. fejezetben található.

3.2 Szórendi és nyelvtani hasonlóság

A mondatok hasonlóságának vizsgálatakor azonban nemcsak a mondatot alkotó szavak egyezősége, hasonlósága a fontos, hanem azok mondatban betöltött szerepe is. Ezért a szemantikai hasonlóságon túl egy másik hasonlósági mértéket is szükséges definiálni. A szövegösszehasonlítás módszertanához forrásként használt cikkben (Li és mtsai, 2006) a szerzők ezt egy szórendi hasonlósággal jellemzik. A szórendi hasonlóság esetében a mondatokat leíró szórendi vektor azt jelöli, hogy a két összehasonlítandó szövegrész szavai vagy elemzési egységei a mondatban hányadik helyen álló szóhoz hasonlítanak a legjobban. Hiszen nem mindegy, hogy „*Egy férfi látott egy vonatot*.” vagy „*Egy férfit láttunk a vonaton*.”. A 3. táblázatban látható, hogy ezen két mondat csupán a szórend figyelembevétele alapján teljesen ugyanaz magyar nyelven. Angol nyelven a szórend pontosan kifejezi a nyelvtani különbséget, de magyar nyelven a kötetlen szórend miatt a nyelvtani szerep figyelembevétele nélkül teljesen azonosnak számítana a két mondat, hiszen a bennük szereplő szavak szótövezett alakja azonos, azonban a mondat jelentése teljesen különböző.



3. ábra: A szóhasonlóság úthossz és mélység függése

3. táblázat: Példa egy szórendi vektorra

		férfi	lát	vonat	
A	Egy férfit láttunk a vonaton.	[1	2	3	]
B	Egy férfi látott egy vonatot.	[1	2	3	]

A szórendi vektorokból a szórendi hasonlóság a következőképp számítandó:

$$S_o(o_1, o_2) = 1 - \frac{o_1 - o_2}{o_1 + o_2} \quad \textbf{(4)}$$

A 4-es képletben o_1 és o_2 a mondatok szórendi vektorait jelölik. Két mondat hasonlósága végül a szemantikai és a szórendi hasonlóság súlyozott átlagaként számítható ki:

$$S(S_1, S_2) = \delta S_s + (1 - \delta) S_o \quad \textbf{(5)}$$

Az 5-ös képletben S_s a szemantikai hasonlóságot, S_o pedig a szórendi hasonlóságot jelöli, δ pedig egy arányt mutat, amelynek értéke 0 és 1 között lehet, Li és mtsai tanulmányukban a 0.85-ös arányt javasolják, és ez jelen modellben is megfelelőnek bizonyult, ezért ez nem került változtatásra.

A 3. táblázat alapján azt a következtetést vonhatjuk le, hogy a szórendi vektor módosításának van létjogosultsága, hiszen a magyar nyelv kötetlen szórendje nem feleltethető meg az angol nyelv kötött szórendjének.

Ezért a szemantikai hasonlóságon kívül kétféle hasonlóságot is lehetne definiálni: szórendi (Li és mtsai, 2006) és nyelvtani hasonlóságot.

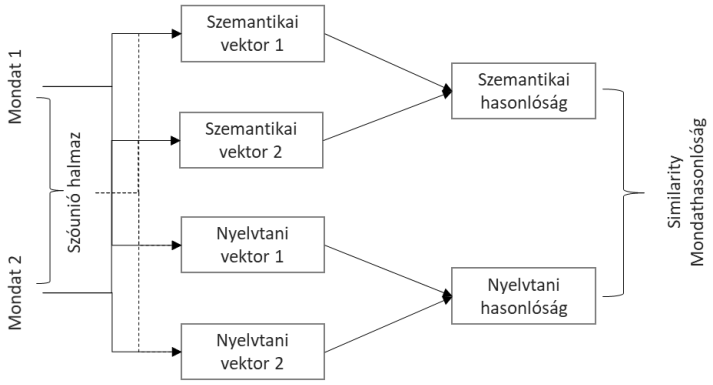
A nyelvtani hasonlóság kiszámításához egy nyelvtani vektor létrehozása látszott megfelelőnek, amely azt jelöli, hogy milyen nyelvtani szerepet töltenek be a két mondatban előforduló szavak. E szerint a magyar nyelvben a mondatbeli szerepeket kell rangsorolni az angol kötött szórendhez mérten, amelyben alapvetően az alanyt követi a tárgy, az állítmány majd ezt követően állnak a határozók. Ezzel a módosítással a 4. táblázatban szereplő vektor értékei a következőképpen módosulnak, és a 3. táblázatban szereplő két példamondat a nyelvtani vektor alapján már nem mondható azonosnak:

4. táblázat: Példa nyelvtani vektorra

		férfi	lát	vonat	
A	Egy férfit láttunk a vonaton.	[3	2	5	]
B	Egy férfi látott egy vonatot.	[1	2	3	]

A tesztelés során felmerült, hogy a szórendi vagy a nyelvtani hasonlóság figyelembevételével kapunk-e pontosabb eredményt, és a nyelvtani vektor használatával bizonyult jobbnak a modell – magyar nyelven.

Összefoglalva tehát két mondat hasonlósága két hasonlósági mérték – a szemantikai és a nyelvtani hasonlóság – súlyozott átlagaként számítható ki, ahogy ezt a 4. ábra is szemlélteti. A hasonlósági érték egy 0 és 1 közötti érték. Két mondat hasonlónak tekinthető, ha egy küszöbszint feletti a hasonlóság értéke. E küszöbszint meghatározása a tesztelés során történt meg, de fontos hangsúlyozni, hogy ez a küszöb minden kutatásnál eltérő lehet, és hasonlóan a módszertan alapjául szolgáló cikkhez (Li és mtsai,, 2006), ennek a küszöbnek az értéke itt sem kerül közlésre.



4. ábra: A szemantikai hasonlóság kiszámításának alapfolyamata

3.3 A módszer magyar nyelven való implementálása

A 3.2 fejezetben a módszertan alapjaiban történtek változtatások, melyek a magyar nyelv struktúrájához mérten alakultak ki. A módszer magyar nyelven való implementálásakor azonban számos további akadályba ütközhet a kutató. A módszertan alapjául

szolgáló cikk (Li és mtsai, 2006) szerzője azonban a tanulmányban semmilyen iránymutatást nem ad a módszer nyelvfüggő jellemzőit illetően, ezért a következőkben ezen kritikus pontokra fogok rátérni, amely nem kifejezetten matematikai vagy statisztikai, hanem inkább nyelvi eredetű.

Egy szövegelemzés szerves része a szövegtisztítás is, és ennek minősége döntő lehet a modell végső teljesítményét illetően.

A szövegösszehasonlító algoritmus Python programozási nyelven került implementálásra, amely esetén eddigi munkám során a magyar nyelven beépített szótövező algoritmusok egyike sem mutatott hatékony eredményeket (nltk package). A szótövezés során a Szegedi Tudományegyetem magyarlánc nevű JAVA alkalmazását használtam. Akadtak azonban olyan esetek, amiket a szótövezésen felül és azzal egyidőben kellett kezelni: a tagmondatra bontást, a tagadás kezelését, jelentésmódosító szavak helyzetét, melléknévi igenevek kezelését és a stopszavazást.

3.3.1 A tagmondatra bontás

A szövegek összehasonlításánál az összehasonlítandó szövegek hosszát is figyelembe kell venni, ezért célszerűnek tartható a szöveg hosszának normalizálása, ami jelen esetben a tagmondatra bontással tűnt kivitelezhetőnek.

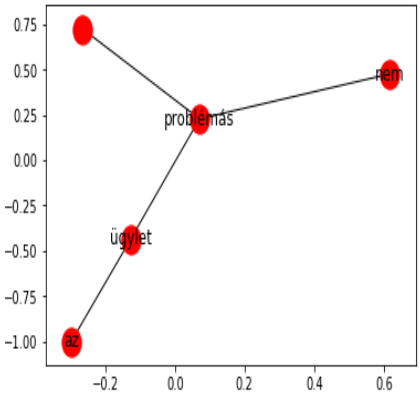
A tagmondatra bontást kétféleképpen közelítettem meg: egyrészről az írásjelek és a kötőszavak felől, másrészről a magyarlánc nyelvtani elemzéséből kiindulva. A magyarlánc alkalmazásban mellérendelő tagmondatok esetében tapasztalhatók hibák, de a nyelvtani elemzése ennek ellenére is hatékonynak bizonyult a tagmondatokra bontás elvégzésében.

Például a következő mondat esetében – *„Az ügyfél problémamentes, megbízható, ezért javasoljuk a pozitív elbírálást.”*. Az írásjelek és kötőszavak felől megközelítve ez a mondat három tagmondatból áll, viszont a magyarlánc alkalmazás szerint ez csupán két tagmondat, mivel az alanyi állítmányokat nem értelmezi külön tagmondatnak. Végül azonban mivel a felsorolásokat hatékonyan tudja értelmezni, ezért ezen szempontot erősebbnek érezve a nyelvi függőségi viszonyok szerinti részekre bontásra esett a választás.

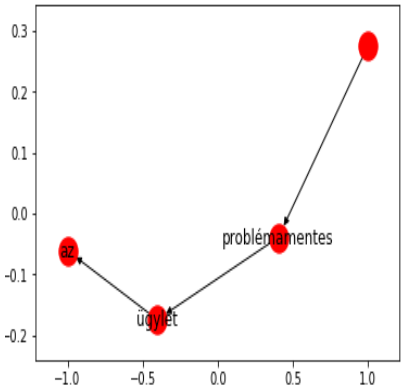
3.2.2. A tagadás kezelése

A módszertan alapjául szolgáló tanulmányban (Li és mtsai, 2006) a szerzők nem térnek ki arra, hogy mi a helyzet azzal, ha két mondat ellentétes jelentésű, hiszen a szemantikai hasonlóság értékkészletét 0 és 1 között határozzák meg. Azonban két mondat nagyon hasonló lehet szókészletében, de már csupán egy tagadás is megváltoztathatja a mondat értelmét. Például nem mindegy, hogy *„Egy banknál sem vezet számlát.”* vagy *„Egy banknál vezet számlát.”*. A tagadószavak kezelése viszonylag egyszerűnek tűnik, azonban többféleképpen is kifejezhetjük azt. Alapvetően a *„nem, nincs, sem, nincsen, nélkül, nélküli”* kifejezések egyszerűen kapcsolódhatnak az adott mondatrészhez, és ez a nyelvtani függőségekből jól látszik is, ezért ezt hozzá tudjuk kapcsolni az adott szóhoz (szavakhoz) automatikusan. Azonban a tagadás nemcsak különálló részként kapcsolódhat a szóhoz, hanem azzal egybeírva is. Példaként említhető az a két mondat, hogy *„Az*

ügylet problémamentes." és az *„Az ügylet nem problémás."*, melyekben a szavak egymáshoz viszonyított kapcsolatát az 5. és a 6. ábra mutatja. Az 5. ábrán látható, hogy a *problémás* szóhoz tartozik egy tagadószó *(nem)*, viszont a 6. ábra szerint a nyelvtani elemzés kimenetében semmi nem utal a tagadásra, pedig valójában a *problémamentes* is ugyanazt jelenti, mint az, hogy nem problémás.



5. ábra: Az ügylet nem problémás nyelvtani kapcsolatai



6. ábra: Az ügylet problémentes nyelvtani kapcsolatai

Azonban nem számítható egyszerűen tagadásnak az összes olyan melléknév, amely *„mentes"* szóval végződik, hiszen elképzelhető olyan eset, hogy például a *problémamentes* és a *jó* összehasonlításánál tévedünk azzal, ha tagadásnak vesszük a problémamentes szót. A tagadás tehát úgy értelmezhető, hogy a jelentés hasonló, de a kapcsolat iránya ellentétes, tehát -1-gyel való szorzás alkalmazható.

Összességében a tagadást több szinten érdemes kezelni, de itt is figyelembe kell venni a nyelvi sajátosságokat és a rendelkezésre álló nyelvi eszköztárt is.

3.2.3 Jelentésmódosított szavak kezelése

A szemantikai hasonlóság megállapításakor felmerül az a kérdés, hogy a bizonyos szófajokból képzett szavakat mennyire vegyük hasonlónak vagy azonosnak ahhoz a szóhoz, amiből képeztük. Jelen esetben például a melléknévi igenevek azzal az igével egyenrangúnak vehetők, amiből képeztük, vagy akár a főnevek -i képzős formái is azonosnak tekinthetők azzal a főnévvel, amiből létrejöttek. Ez a tulajdonság a magyarlánc a nyelvtani elemzésből látszik, ezért automatikusan ezeket szabályszerűen ki lehet szűrni, azonban vannak kivételek is, amelyek felvétele manuálisan történhet (például -i képzős melléknevek, amik főnévből képződtek, vagy melléknévi igenevek).

3.2.4 Stopszavazás

Minden szöveganalitikai elemzés során elképzelhető olyan helyzet, hogy bizonyos szavak nem relevánsak a kutatási kérdés szempontjából, és a tisztítási folyamat során ezeket el kell távolítanunk azért, hogy hatékonyabban tudjon működni a modell. Jelen esetben a vektorok hosszának növekedése csökkentené a hasonlóság mértékét, és nagyobb hiba is felmerülhetne, ezért a stopszavak meghatározását mindig nyelv- és kutatásfüggően kell meghatározni és nem szabad csak és kizárólag a beépített stopszólistákra hagyatkozni, hanem ez egy nagyon fontos szakértői döntés kell legyen.

4 Programkód részletek

#koszinusz távolság

```
def cosine_similarity(x,y):
    numerator = sum(a*b for a,b in zip(x,y))
    denominator = square_rooted(x)*square_rooted(y)
    return round(numerator/float(denominator),5)
```

#szemantikai hasonlóság

```
def semantic_similarity(word1, word2, G, a=0.2, b=0.45):
    if word1 == word2: #ha a két szó teljesen azonos
        return 1
    else:
        try:
            l=shortest_path_length(G, word1,
word2)["length"] #legrövidebb út hossza
            h=shortest_path_length(G, word1,
word2)["depth"] #közös ős mélysége a hálózatban
            return (math.exp(-1*a*l)*((math.exp(b*h)-
math.exp(-1*b*h))/(math.exp(b*h)+math.exp(-1*b*h))))
        except TypeError:
                return (np.nan)
```

5 Eredmények

A módszertan alapjául szolgáló tanulmány (Li és mtsai, 2006) magyar nyelven való implementálása nem volt kihívásoktól mentes. A 3. fejezetben leírt módosítások segítségével azonban egy hatékonyan működő szövegösszehasonlító rendszer jött létre. Két egymástól független teszten a 7. ábrán szereplő eredmények születtek a módszerek alkalmazásával (100-100 hasonló, és 150-150 nem hasonló mondatpár):

similar_dummy

similarity_dummy		SIMILAR	not_SIMIL...
not_similar	Count	37	139
	Column %	36.275	92.667
similar	Count	65	11
	Column %	63.725	7.333

similar_dummy

similarity_dummy		SIMILAR	not_SIMIL...
not_similar	Count	40	136
	Column %	40.000	90.667
similar	Count	60	14
	Column %	60.000	9.333

7. ábra: Modell teljesítménye

Az értékelésnél az első- és másodfajú hibát tekintve súlyosabb hibának számított az, hogy ha egy mondatpárról a modell azt mondta, hogy hasonlóak, miközben jelentésükben teljesen különbözőek voltak. A modell szerinti álpozitív értékek sokkal súlyosabb hibának számítottak, mint az álnegatívak, de összességében a 60 %-os találati arány és a maximum 10 %-os másodfajú hiba még megfelelő teljesítménynek mondható.

6 Konklúzió

Összességében elmondható tehát, hogy a nyílt forráskódú programnyelvek terjedésével egyre több ötlet és lehetőség kerül nyilvánosságra a természetesnyelv-feldolgozás területén, amelynek az implementálása sokszor szükségessé válhat egyéb idegen nyelven. A kutatási kérdésünknek megfelelő eszköztár és módszertan megtalálása nem egyszerű feladat, és jelen tanulmány azt hivatott bemutatni, hogy egy jó kiindulási alapból lehet fejleszteni, de nem határok nélkül. A természetesnyelv-feldolgozó rendszerek esetén figyelembe kell venni az adott nyelv jellegzetességeit, szókészletét és a nyelv elterjedtségét is, és a rendelkezésre álló elemzési eszköztárt, és a munka nagy részét nem csupán egy más forrásból származó ötlet lemásolása és reprodukálása jelenti. Az implementálás mögött hosszas szakértői munka van, mely segítségével létrehozható az a témaspecifikus tudás, amely nélkül önmagában egy idegen forrásból származó programkód nem tud lefutni. Úgy gondolom, hogy ez a határa a mai természetesnyelv-feldolgozás térhódításának, ugyanakkor ebben a témaspecifikus szakértői tudásban rejlik a természetesnyelv-feldolgozás továbbfejlesztése is.

Jelen tanulmány iránymutatásai a magyar nyelvű szöveganalitika egy részébe engednek betekintést, de korántsem tekinthetők teljes leírásnak. A magyar nyelvű témaspecifikus források repertoárjának bővítése hozzájárulhat a kutatási terület fejlődéséhez, és ennek megfelelő mértékű növekedése esetén könnyebbé válhat az idegennyelvű módszertanok magyar nyelven való implementálása. Azonban a nyelvek jellegzetessé-

geinek mindig lesznek olyan pontjai, amelyeket figyelembe kell venni egy ehhez hasonló feladat esetén, és ezen szempontok megtalálása a további hasonló témájú kutatások létjogosultságát adhatja.

Hivatkozások

Lapata, M., Barzilay, R. Automatic evaluation of text coherence: Models and representations. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. (2005)

Li, Y., McLean, D., Bandar, Z., O'Shea, J., Crockett, K.: Sentence similarity using semantic nets and corpus statistics. In: IEEE Transactions on Knowledge and Data Engineering. 18. pp. 1138-1150. (2006)

Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1. pp. 775-780. Boston, Massachussets (2006)

Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of the Fourth Global WordNet Conference GWC, pp. 310-320. (2008)

Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Information Processing and Management 2(32). pp. 193-207.(1997)

Slimani, T.: Description and Evaluation of Semantic Similarity Measures Approaches. International Journal of Computer Applications 80.10 pp. 25–33. (2013)

Bu-Bor-éK: grafikus címkenormalizáló eszköz

Novák Attila[1,2], Novák Borbála[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat A webes portálokon megjelenő tartalmakat gyakran tematikus címkékkel látják el, amelyeket jelenléte többek között hatékonyabb kereshetőséget, a webes keresőkben jobb találatokat eredményez, illetve a kapcsolódó vagy személyre szabott tartalmak megjelenítéséhez is használható. A kulcsszavakat gyakran manuálisan és nem egységesen rendelik a tartalmakhoz, ez gyakran a címkekészlet nemkívánatos elburjánzásához vezet. Cikkünkben egy olyan grafikus eszközt mutatunk be, amelyet az említett probléma kezelésére címkebeágyazási modellek kétdimenziós megjelenítéséből kiindulva többek között címkekészletek normalizálására és szerkesztésére lehet használni. A címkekészlet szerkesztésére szolgáló eszközben a vizuális modell bejárható, a címkék kereshetőek, szerkeszthetőek, címkeosztályokba sorolhatóak, a szinonim címkék összevonhatóak.
Kulcsszavak: információkinyerés, annotáció, lexikai erőforrások, kulcsszónormalizálás, grafikus eszköz

1. Bevezetés

Az utóbbi években lezajlott paradigmaváltás eredményeképpen mára nem túlzás azt állítani, hogy a nyelvtechnológiában előforduló szinte minden feladatra neurális hálózatok alkalmazásán alapuló megoldásokkal érhető el a legjobb eredmény. Míg kezdetben a szóbeágyazási modellek önmagukban is lenyűgöző szemantikai reprezentációt produkáltak, addig mára a világ élmezőnyébe tartozó kutatói egyre bonyolultabb architektúrákat alkalmaznak egy-egy feladat megoldására. Ezeknek az összetett hálózatoknak a belső működése sok esetben már teljesen értelmezhetetlen. Az azonban még mindig igaz, hogy a neurális modellekben a szavak, illetve egyre inkább a szavaknál kisebb lexikai egységek nem szimbolikus formában, hanem néhányszáz dimenziós vektorokként jelennek meg.

Az általában köztes reprezentációként, de akár végeredményként létrejövő vagy éppen egy hálózat bemeneteként szolgáló sokdimenziós vektorok értelmezése és azok minőségének ellenőrzése nehéz feladat. A szakirodalomban elterjedt közvetlen kiértékelési módszerek a hasonlósági listák, illetve analógiák vizsgálatával ellenőrzik a szóbeágyazások minőségét (l. pl. Faruqui és mtsai (2016), Schnabel és mtsai (2015)), vagy valamilyen a beágyazási modellt egy ráépülő feladat megoldásához használó komplexebb modell teljesítményének változásán

keresztül próbálják közvetetten jellemezni a beágyazási modellek minőségét. Egy másik megközelítés a sokdimenziós vektorok terét két-három dimenzióba képezi le, ami már könnyen vizualizálható, így ránézésre is áttekinthetővé teszi a modellben szereplő elemek reprezentációját, azok egymáshoz képesti elhelyezkedését a modell alkotta térben. Ez utóbbi módszer kvantitatív kiértékelésre kevésbé alkalmas, viszont nagyobb rálátást, jobban áttekinthető megjelenítést tesz lehetővé.

Ezek a kiértékelésre és elemzésre szolgáló módszerek azonban statikusak, a modellben létrejött reprezentációnak csupán a megjelenítésére szolgálnak. Ebben a cikkben egy olyan eszközt mutatunk be, amely a többszáz dimenziós beágyazások kétdimenziós leképezéséből kiindulva lehetővé teszi a megjelenített elemek mozgatását, szerkesztését és összevonását. Az eszköz hatékonyan alkalmazható többek között zajos címkekészletek kézi tisztítására beágyazásalapú címkemodellből kiindulva. A tisztított címkekészlettel a modell újrataníthatóvá, és pontosabb, egységesebb eredményt adó modell nyerhető.

2. Motiváció

A bemutatásra kerülő eszközt két motivációs példán mutatjuk be. Az első példában egy szövegcímkéző rendszer tanításakor használt címkekészlet normalizálása a feladat, a másodikban pedig egy a szavakhoz szemantikai osztályokat rendelő rendszer osztályrendszerének átalakítása.

2.1. Szövegcímkézés

A webes hírportálokon megjelenő szövegeket gyakran különböző tematikus címkékkel látják el, melyek lehetővé teszik a látogatók számára, hogy kifejezetten valamilyen számukra érdekes témával, személlyel, eszközzel stb. kapcsolatos cikkeket vagy egyéb tartalmakat jelenítsék meg. Másrészt a kulcsszavakat az adott cikkhez kapcsolódó egyéb cikkek vagy tartalmak megjelenítéséhez is használják, illetve szerepet játszanak a címkék a keresőmotorok (pl. a Google) találatrangsorolási algoritmusaiban is.

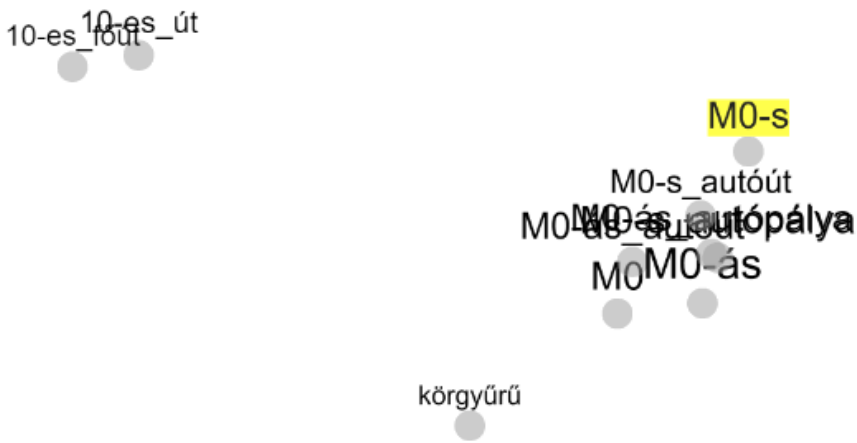
Egy szöveghez az annak tartalmához kapcsolódó tematikus kulcsszavak automatikus hozzárendelésére számos algoritmikus megoldás létezik. Egy ilyen kifejezetten magyar nyelvű sajtószövegek címkézésére szolgáló eszközt mutat be például Farkas (2009). Ennek ellenére sok online is megjelenő szövegarchívumban a cikkekhez a szerző/szerkesztő által egyedileg kézzel hozzárendelt kulcsszavak szerepelnek (pl. Farkas (2009) sem említi, hogy az ott bemutatott algoritmust újonnan születő cikkek címkézésére (vagy annak segítésére használták volna). A kézi címkézést néha erre szakosodott (általában könyvtáros végzettségű) szakember végzi, azonban sokszor inkább maguk a szerzők végzik el ezt a feladatot is.

Ebből kifolyólag az egy archívumon belül használt címkekészlet gyakran nem egységes, a szerzők ugyanannak a címkének különböző (gyakran elírt) formáit használhatják: *M0-ás autópálya, M0-ás, M0-s autópálya, M0-s, M0-ás autóút,*

M0, M0-s autóút. Bár a kézzel címkézett szövegek jól használhatóak egy automatikus címkéző rendszer tanításához, a kulcsszavak változatossága miatt a rendszer mért pontossága alacsonyabb lesz az elvártnál.

Egy folyamatban levő projekt keretében sajtószövegek automatikus címkézésére vállalkoztunk, amelyre a fastText programcsomag (Joulin és mtsai, 2017) címkézőalgoritmusát használjuk. Az osztályozóhálózat(ok) bemenetén az adott szöveg tokenjeinek, illetve token-n-eseinek reprezentációja jelenik meg (a bennük szereplő különböző hosszú karakter-n-gramok reprezentációjának átlagaként), és az osztályozó ehhez a szövegreprezentációhoz és az egyes lehetséges címkékhez rendel illeszkedési értéket multinomiális logisztikus regresszió alkalmazásával. Megfelelő küszöbérték választása mellett az adott szövegre jól illeszkedő kulcsszavak elválaszthatóak a kevésbé jól illeszkedőktől. Bár megjelenése óta a fastText modellnél jobban teljesítő szövegosztályozó modellek is készültek (a cikk írásának idején az ilyen jellegű feladatokban a mélyneurális XLNet architektúra nyújtja a legjobb teljesítményt több angol nyelvű adatbázison (Yang és mtsai, 2019)), ezeknek komplexitása, hardver- és futásiidő-igénye a pontosságbeli teljesítménykülönbséget jóval meghaladó mértékben nagyobb, mint a fastTexté.

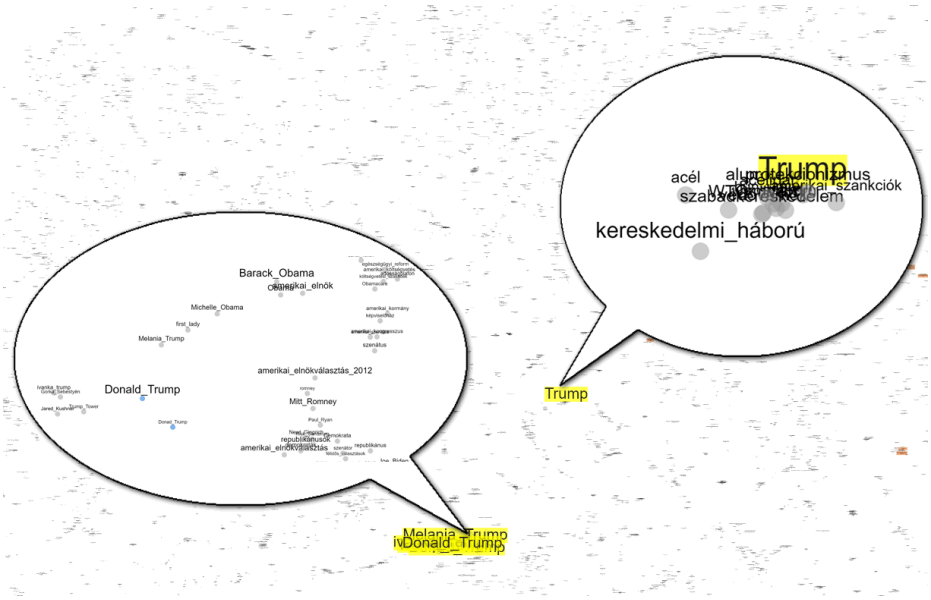
A betanított modell címketerében megfigyelhető, hogy egy címke különböző írásváltozatainak a reprezentációja a beágyazási térben egymáshoz közel helyezkedik el, mert hasonló témájú cikkeket címkéznek ugyanannak a kulcsszónak a különböző változataival (1. ábra).



1. ábra: Az M0-s címke írásváltozatainak elhelyezkedése a címketérben

Időnként megfigyelhetőek eltérések ettől az alapvető mintázattól, de ennek mindig a szövegkorpuszra, a címkehasználat egyedi sajátosságaira, illetve a címkék többértelműségére visszavezethető magyarázata van. Modellünk például egy-

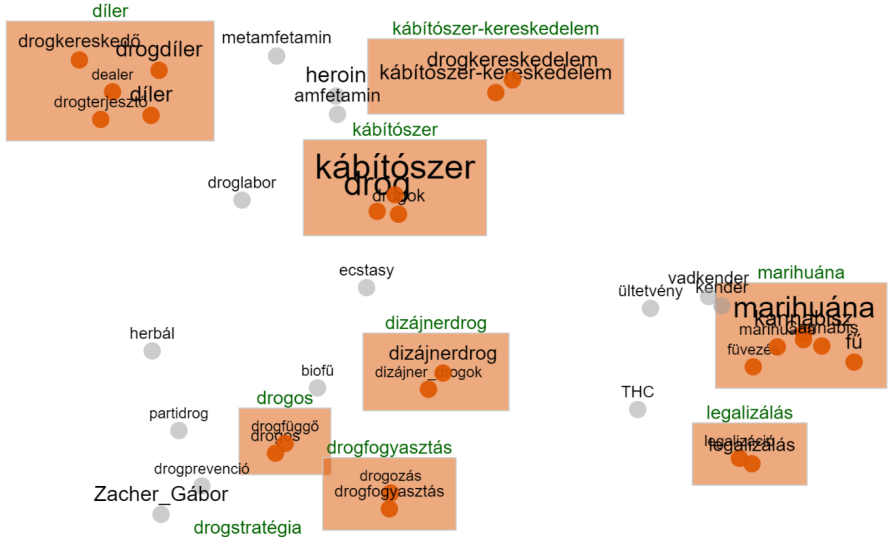
értelműen megragadta azt a sajátosságot, hogy az adott korpuszban a gazdasági témájú cikkek szerzői az amerikai elnököt következetesen *Trump*-nak címkézik, míg a politikai cikkek szerzői a keresztnevét is használva (azt időnként elírva) *Donald (Donad, Donal) Trump*-nak. Így a keresztnév nélküli Trump címke a keresztnevesektől viszonylag távol a kereskedelmi háborúval kapcsolatos címkék között szerepel (2. ábra). Többértelműségből fakadóan került pl. a *magyar csapat* címke viszonylag távol a sporttól (annak ellenére, hogy olimpiai témájú cikkek is viselik ezt a címkét) és közel Németh Szilárd rezsibiztoshoz a 2014 eleji politikai jellegű „Magyar Csapat"-kezdeményezésről szóló cikkek hatására. Számos esetben a szinonim címkék nemcsak különböző írásváltozatokat, hanem lényegében ugyanannak a fogalomnak különböző eredetű/stílusú megnevezéseit ölelik fel, pl. *fű, marihuána, kannabisz* stb. (3. ábra).



2. ábra: Trump címkéi egymástól távol

2.2. Szemantikai osztályozás

A másik feladatban a Dologfelismerő (Novák és Siklósi, 2017) által használt címkerendszer normalizálása volt a cél. A Dologfelismerő létező szemantikai erőforrásokból (Roget's Thesaurus (Chapman, 1977), Longman (Summers, 2005), 4lang (Kornai és mtsai, 2015)) szóbeágyazások segítségével készített modell alapján rendel szemantikai kategóriákat, illetve tulajdonságokat bármilyen (magyar vagy angol) szóhoz. A Dologfelismerő létrehozásakor is ismert probléma volt a



3. ábra: fű, marihuána, kannabisz

felhasznált erőforrások címkekészletének régiessége (Roget's Thesaurus), illetve a különböző erőforrások címkekészletének összehangolása. Bár a Dologfelismerő sokszor jól meg tudja ragadni egy szó szemantikai tulajdonságait és kategóriáit, a megjelenített címkékből ez nem mindig látszik. Az értelmetlennek tűnő címke reprezentációjának létrehozásához használt szavak csoportját vizsgálva sok esetben azonban kiderül, hogy csupán az eredeti erőforrásokból átvett megnevezés a téves. Erre korábban az eredeti címkék klaszterezéssel történő felbontása volt a megoldás. Például a *Combatant* kategóriába tartozó szavak közül a *charger, battery, file, monitor* külön klaszterbe került, hiszen ezek ma már inkább számítástechnikai/elektronikai jelentést hordoznak. Így bár maga a kategóriacímke nem feltétlenül jellemzi jól a hozzá tartozó szemantikai jegyet, de a klaszterezés során hozzáadott numerikus index alapján azonosítható és jól elválasztható ez a kategória a *Combatant* címkéhez tartozó szavakból létrejött többi, katonai kifejezéseket tartalmazó kategóriától. Ez a megoldás azonban csupán egy technikai megoldás volt, a címkék elnevezése továbbra sem feltétlenül tükrözte az általuk reprezentált tartalmat.

Célunk volt a Dologfelismerő címkekészletének normalizálása, de mivel a modell címkekészletének mérete miatt a címkék kézzel való átnézése nem könnyű feladat, ráadásul sok esetben nemcsak a címkék átnevezése, hanem több címke összevonása is szükségesnek tűnt.

3. A címkekészletek normalizálására szolgáló eszköz

A tipikusan néhány száz dimenziós címkebeágyazási modellben[1] megjelenő címkereprezentációkat a vizualizációhoz és a grafikus szerkesztéshez először két dimenzióba vetítjük. Bár a tavalyi MSZNY-en bemutatott beszédfelismerők vizualizációjával kapcsolatos eredményeken (Grósz és Tóth, 2019), és különösen az egyik szerzővel folytatott későbbi beszélgetésen felbuzdulva kísérleteztünk autoencoder alapú vizualizációval is, az eredmények a mi esetünkben nem bizonyultak használhatónak, így a lokális kapcsolatokat jobban megőrző klasszikus t-SNE vizualizációs algoritmus (van der Maaten és Hinton, 2008) alkalmazása mellett maradtunk.

A javascript-alapú cytoscape.js gráfvizualizációs és -szerkesztő csomag (Franz és mtsai, 2015) felhasználásával készítettük el a címketér elemeinek szerkesztésére, illetve az azonos szerepű címkék összevonására szolgáló testre szabott böngészőalapú szerkesztőeszközünket.

A címketér t-SNE algoritmussal kapott 2 dimenziós képét[2] a Cytoscape-pel kompatibilis json formátumba konvertáljuk, és ehhez hozzáadjuk a gyakorisági adatokat (illetve a megjelenítéshez a csomópontoknak a gyakoriság logaritmusával arányos méretét). A megjelenített címketérkép egérrel/trackpaddel navigálható, zoomolható, az egyes címkék mozgathatóak.

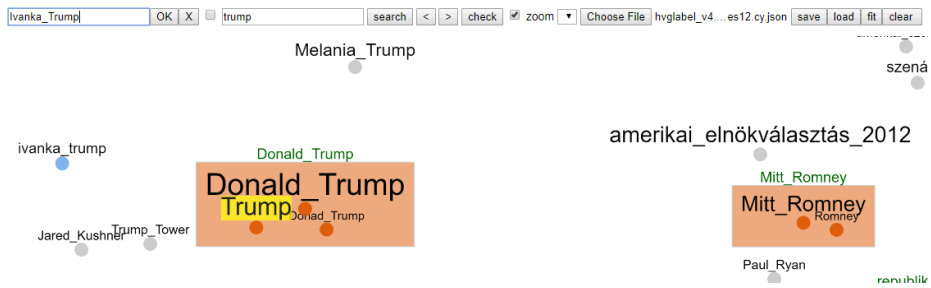
A címkéket egymás közelébe mozgatva vagy az ekvivalens címkék kijelölése után a megfelelő billentyű-egérgombkombináció megnyomásával azok szinonimacsoportba csoportosíthatóak. A csoportot reprezentáló narancsszínű téglalapként megjelenő szülőcsomópont eredő címkéje automatikusan a csoportban szereplő leggyakoribb címke lesz (4. ábra: *Donald Trump, Mitt Romney*), de más címke is kiválasztható, illetve a címkék szerkeszthetőek. Szerkesztéskor, illetve összevonáskor mindig megfelelően nyomon követjük az eredeti címkéket is, hiszen az eszköz célja éppen az, hogy az eredeti túlságosan változatos, illetve hibás címkéket az adatbázisban javítani, illetve egységesíteni tudjuk.

A megjelenítésre és szerkesztésre szolgáló felület felett helyeztük el a címkék szerkesztésére, a keresésre, a modell betöltésére és elmentésére és különböző statisztikai információk megjelenítésére szolgáló vezérlőelemeket (4. ábra fölül). Lehetőség van a szerkesztendő/javítandó címke egy billentyűlenyomásra történő automatikus kis/nagybetűsítésére is. Erre viszonylag gyakran van szükség a hibásan csupa kisbetűvel írt nevek miatt (4. ábra: itt éppen az *ivanka trump* címke nagybetűsítése történik a felül baloldalt látható címkeszerkesztő mezőben).

Az eszköz lehetőséget ad arra, hogy címkékre és címkerészletekre keressünk. Ilyenkor az illeszkedő címkéket (sárgával) kiemelve és kinagyítva jeleníti meg az eszköz (2. ábra), illetve lehetőség van csak az illeszkedő címkéket magába foglaló területre való automatikus ráközelítésre is. A kinagyított/kiemelt címkéket

[1] A cikkben említett tematikuscímke-beágyazási modell dimenziószáma 100, a Dologfelismerő modellé 300.

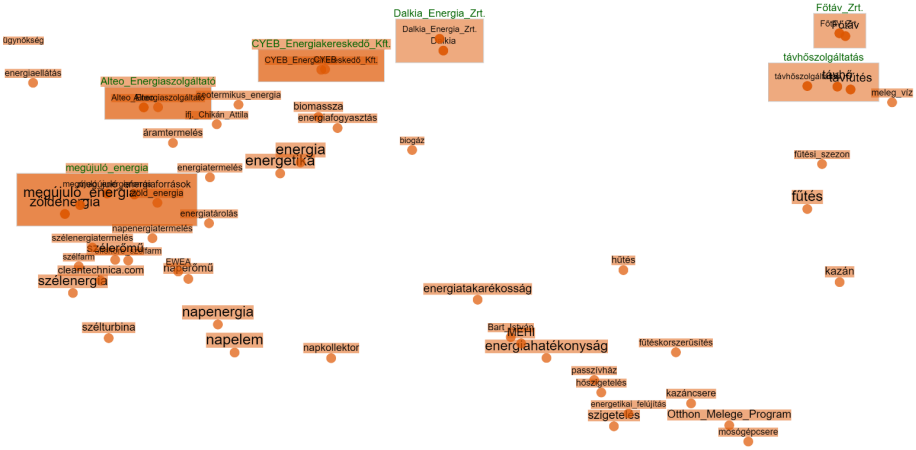
[2] A t-SNE perplexitásparamétereként 50-es értéket használtunk. Ha a megjelenítendő elemszám kisebb, mint a perplexitásparaméter háromszorosa, akkor a preplexitásértéket a megjelenítendő elemek harmadára állítjuk be.



4. ábra: A grafikus szerkesztő vezérlőelemei – Címke automatikus nagybetűsítése szerkesztéskor (*ivanka trump*)– Szinonimacsoportba rendezett címkék az eredő címkéjükkel (*Donald Trump, Mitt Romney*) – Trump címkéi itt már együtt

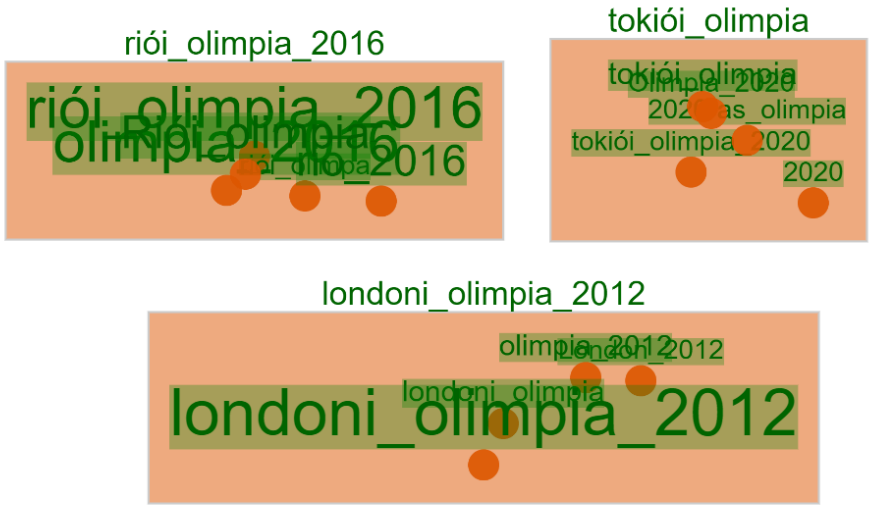
abban az esetben is könnyen a megfelelő helyre lehet mozgatni, ha azok az eredeti térben esetleg a velük ekvivalens címkéktől messzebb találhatóak (mint a korábban említett *Trump* címke esetében). Lehetőség van az illeszkedő címkéken való végiglépkedésre is mindig automatikusan az épp soron következő címke környezetére ráközelítve (l. 4. ábra: *Trump* sárgával kiemelve).

A már átnézett nem összevonandó címkéket megjelölhetjük ellenőrzöttként. Ezek címkéje az összevont kulcsszavakat reprezentáló dobozokhoz hasonló narancsszínű hátteret kap (5. ábra). Így adminisztrálhatjuk a munka előrehaladását.



5. ábra: Az átnézett, kijavított címkéket készként megjelölhetjük.

Az eszköz lehetővé teszi speciális címkeosztályok kezelését is. Pl. a tematikus címkézésnél megkülönböztethetünk olyan címkéket, amelyek egy-egy időben jól körülhatárolt eseményt jelölnek (pl. egy konkrét sportverseny, fesztivál, kiállítás, díjátadó, baleset vagy választás). Ezek reprezentációja nagyon hasonlít bármelyik másik hasonló eseményéhez (pl. a 2018-as Oscar-gála legjobban a többi Oscar-gálához hasonlít), azonban ezek hosszú távon a címkézőrendszer szempontjából valószínűleg nem hasznos címkék. A szerkesztő lehetővé teszi az ilyen címkék megjelölését, és a hosszú távon őket helyettesítő általános címkékhez kapcsolását.



6. ábra: Az egyedi eseményeket jelölő címkék megjelölése.

Lehetőség van a címketérkép aktuális állapotával kapcsolatos statisztikai adatok megtekintésére is (feldolgozottnak jelölt, még feldolgozásra váró, átnevezett, speciálisnak jelölt (pl. egyedi eseményeket jelölő), illetve az összevont, valamint az összevontak fölé rendelt (szülő-) címkék száma).

A Dologfelismerő címkéinek szerkesztéséhez kiegészítettük az eszközt egy olyan funkcióval, ami lehetővé teszi a címkéhez kapcsolódó példaszavak megjelenítését (egyszerűen a címke kiválasztásával), ami alapján egyrészt a címke által jelölt halmaz szemantikai koherenciája felmérhető (és a címke megjelölhető, ha a szóhalmaz nem koherens), másrészt a címke a halmazt ténylegesen fedő fogalomra átnevezhető. Illetve ugyanez a funkció segíti a címkék összevonását is. A címketér alapján való megjelenítés és a konkrét példaszavak címkék alá rendelése azt is lehetővé teszi, hogy – ellentétben magával a szóbeágyazási térrel, ahol a

többértelmű szavak nem jelennek meg több példányban – a konkrét példaszavak több különböző címke alatt megjelenhetnek, akár különböző értelemben.

Az eszköz használatával olyan a természetes nyelvhasználatból adódó szemantikai csoportok is feltárultak, amelyek egyébként nem merültek volna fel bennünk: pl. heraldikai elemek, szabász-varrászati eljárások, stb. Emellett az egyébként hasznos és az eredeti címkekészletből mindenképp megtartandónak látszó címkéken/csoportokon belül (pl. betegségek) tapasztaltuk olyan jellegű alcsoportok megjelenését, amelyek inkább egy mindennapi 'józan ész' jellegű ontológiára utalnak (pl. a betegségek szétválása veszélyes–nem veszélyes betegségekre). Ezeket az átnevezett címkékben tükröztettük.

4. Összefoglalás

Cikkünkben egy címkekészletek normalizálására és szerkesztésére szolgáló grafikus eszközt mutattunk be. Címkebeágyazási modellből a t-SNE algoritmussal nyert 2D vizualizációból kiindulva lehet egyszerű műveletekkel összevonni a szinonim címkéket, átnevezni a nem megfelelő nevet viselőket, és ily módon jobb minőségű tanítóanyagot hozni létre a nyelvi modellek építéséhez.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Chapman, R.: Roget's International Thesaurus. Harper Colophon Books, Crowell (1977), https://books.google.hu/books?id=9VhQAAAAMAAJ

Farkas, R.: Az origo automatikus címkézési projekt tapasztalatai. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009). pp. 84–92. Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2009)

Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C.: Problems with evaluation of word embeddings using word similarity tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 30–35. Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://www.aclweb.org/anthology/W16-2506

Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sümer, S.O., Bader, G.D.: Cytoscape.js: a graph theory library for visualisation and analysis. In: Bioinformatics (2015)

Grósz, T., Tóth, L.: Mély neuronhálós beszédfelismerők működésének értelmező elemzése. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 287–298. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2019)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017), https://www.aclweb.org/anthology/E17-2068

Kornai, A., Ács, J., Makrai, M., Nemeskey, D.M., Pajkossy, K., Recski, G.: Competence in lexical semantics. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. pp. 165–175. Association for Computational Linguistics, Denver, Colorado (June 2015)

van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)

Novák, A., Siklósi, B.: A Dologfelismerő. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017). pp. 25–36. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2017)

Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015), https://www.aclweb.org/anthology/D15-1036

Summers, D.: Longman Dictionary of Contemporary English. Longman Dictionary of Contemporary English Series, Longman (2005), https://books.google.hu/books?id=4zktAAAACAAJ

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. CoRR abs/1906.08237 (2019), http://arxiv.org/abs/1906.08237

Mély neuronhálós akusztikus modellek súlyinicializálásának vizsgálata

Pintér Ádám[1], Tóth László[1], Gosztolya Gábor[1,2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
{ tothl, ggabor } @ inf.u-szeged.hu

Kivonat Az automatikus beszédfelismerés területén az akusztikus modellezésben gyakorlatilag egyeduralkodókká váltak a mély neurális hálók. Az irodalomban számos megoldást találunk arra, hogy hogyan érdemes beállítani a különböző paramétereket a DNN akusztikus modellek tanítása során, azonban általában kevés figyelmet szentelnek annak, hogy a hálók súlyait hogyan érdemes inicializálni. Eközben a gépi tanulási irodalomban ez egy igen aktív terület; a közelmúltban több stratégia is napvilágot látott a DNN kezdősúlyainak beállítására. Jelen munkánkban három ilyen eljárást tesztelünk mély neurális hálós akusztikus modellekben, három különböző aktivációs függvényt (szigmoid, ReLU és szoftplusz) használva. Eredményeink alapján mindenképp érdemes valamilyen speciális súlyinicializálási eljárást alkalmaznunk, ugyanakkor a három vizsgált stratégia (Glorot, He és Edge of Chaos) használatával elért fonémaszintű hibaarányok között nem találtunk szignifikáns különbséget.
Kulcsszavak: beszédfelismerés, mély neurális hálók, súlyinicializálás, Glorot inicializálás, He inicializálás, Edge of Chaos

1. Bevezetés

Az elmúlt évtizedben a mély neurális hálók (Deep Neural Networks, DNN) nagyon gyorsan elterjedtek a gépi tanulás szinte minden területén. Az automatikus beszédfelismerésben is gyakorlatilag egyeduralkodókká váltak az akusztikus modellezés részfeladatán, mely elsősorban az általuk elérhetővé váló alacsony hibaarányoknak köszönhető. A beszédfelismerési feladatban ugyanakkor számos olyan részprobléma található, melyre valamilyen speciális algoritmus használata terjedt el (pl. kapcsolt állapotok létrehozása, vagy az akusztikus modell felvételszintű annotációira optimalizáló tanítási eljárások), és ezek neurális hálókra adaptálása folyamatosan zajlik (Grósz és mtsai, 2015; Zhu és mtsai, 2015; Grósz és mtsai, 2017). Emellett a hálók tanítása számos új hiperparaméter behangolását és az akusztikumra fókuszáló speciális tanítási technikák vagy metódusok kifejlesztését is magával vonta (ilyen pl. a Connectionist Temporal Classification (Graves és mtsai, 2006)).

Jelen cikkünkben is a DNN-tanítás egy „hiperparaméterére” fókuszálunk: azt vizsgáljuk meg, hogy a mély neurális hálók mennyire érzékenyek a súlyok kezdeti

értékeire. Habár a súlyokat mindig valamely valószínűségi eloszlást követve választjuk véletlenszerűen, az ezen eloszlást meghatározó paraméterek (jellemzően a szórás) kiválasztására számos stratégiát mutattak be az elmúlt években, és általánosságban is igen aktívan kutatott terület (ld. pl. (He és mtsai, 2015; Poole és mtsai, 2016; Schoenholz és mtsai, 2017; Pennington és mtsai, 2017; Hanin és Rolnick, 2018; Pretorius és mtsai, 2018)). Tudomásunk szerint nem született még olyan tanulmány, amely különböző súlyinicializálási eljárások fonéma- vagy szószintű hibaarányait vizsgálta volna az automatikus beszédfelismerés problémakörében. Vizsgálatunk aktualitását növeli, hogy a közelmúltban jelent meg az Edge of Chaos (röviden EOC, (Hayou és mtsai, 2019)) súlyinicializálási eljárás, mely kifejlesztői szerint lehetővé teszi extrém mély neurális hálók tanítását is. Bár jelen tanulmányunkban nem kísérlünk meg ilyen extrém struktúrájú DNN-alapú akusztikus modellt tanítani, egy ilyen súlyinicializáló eljárás akár alacsonyabb hibaarányokhoz is vezethet.

2. Mély neurális hálók súlyinicializálási stratégiái

A neurális hálókat jellemzően egy iteratív hibavisszaterjesztési (backpropagation) eljárással szokás tanítani. Az eljárásról ugyanakkor ismert, hogy nem garantálja a globális optimumot, hanem lokális optimumhoz vezet. Több rejtett réteg esetén (és a hagyományos szigmoid vagy tanh aktivációs függvényeket alkalmazva) ráadásul föllép a „megszűnő gradiens” (vanishing gradient, (Hochreiter és mtsai, 2001)) néven ismert effektus, mely azt eredményezi, hogy a túl nagy értékű súlyok miatt a mélyebben elhelyezkedő rétegek súlyai nem változnak érdemben (azaz a háló nem tanul). Túl kis súlyok esetén pedig, mivel a tanh és szigmoid függvények 0 körül gyakorlatilag lineárisak, elveszítjük a modell nemlinearitását, valamint a gradiensek „elszabadulhatnak” (exploding gradient). Emiatt létfontosságú, hogy a súlyokat a megfelelő intervallumban tartsuk, illetve onnan is indítsuk.

A következőkben részletesebben ismertetünk három eljárást a kezdősúlyok meghatározására. Viszonyítási alapnak azt tekintettük, hogy a súlyokat egy normális vagy egyenletes eloszlásból vettük, 0 átlaggal. Az értékek szórását ekkor előzetes tesztekkel $0,001$-ben határoztuk meg.

2.1. Glorot súlyinicializálási eljárása

A bevezetőjéről elnevezett Glorot-féle (vagy Xavier-féle) stratégia alapötlete, hogy az egyes rétegek kimeneteinek varianciáját azonos értéken tartsa, hogy az ne csökkenjen, ahogy a hiba visszaterjesztése a háló mélyebben fekvő rétegei felé halad (Glorot és Bengio, 2010). Mivel egy teljes kapcsolású (fully connected) hálóban minden neuron kapcsolatban áll az előző és a következő réteg összes neuronjával, a metódus szerint a súlyok szórása a következőképpen alakul:

$$\sigma = \sqrt{\frac{2}{n_{bemenet} + n_{kimenet}}} \tag{1}$$

míg az átlag 0. Amennyiben a súlyokat normális helyett egyenletes eloszlás szerint választjuk, azoknak praktikusan az

$$U = \Big[-\frac{\sqrt{6}}{\sqrt{n_{bemenet} + n_{kimenet}}}; \frac{\sqrt{6}}{\sqrt{n_{bemenet} + n_{kimenet}}} \Big] \tag{2}$$

intervallumból kell jönniük.

2.2. He súlyinicializálási eljárása

Glorot és Bengio cikke idején az elterjedt aktivációs függvények a szigmoid és a tanh függvények voltak, melyek nulla körül szimmetrikusak és deriváltjuk megközelítőleg egy (azaz a függvény lineáris). Ezt kihasználva elhanyagolhatták a levezetésből az aktivációs függvény alkalmazását. Ez a lépés azonban a később elterjedt függvények (pl. ReLU) esetén nyilvánvalóan nem megalapozott. Az előző számítások adaptálását végezték el He és munkatársai (He és mtsai, 2015). Eredményeik alapján normális eloszlás használata esetén 0 átlaggal és az alábbi szórással kell kiválasztanunk a súlyokat:

$$\sigma = \sqrt{\frac{2}{n_{bemenet}}}. \tag{3}$$

Egyenletes eloszlást használva a kezdősúlyok intervalluma a következő lesz:

$$U = \Big[-\sqrt{\frac{6}{n_{bemenet}}}; \sqrt{\frac{6}{n_{bemenet}}} \Big]. \tag{4}$$

2.3. Edge of Chaos

A „Káosz határa” (Edge of Chaos, EOC) inicializálási stratégia más megközelítésen alapszik. Az alapötlet az, hogy egy csupa véletlenül inicializált súlyt tartalmazó, teljesen kapcsolt mély neurális hálón különböző bemeneti értékekre megvizsgálva az előálló kimeneteket elvárjuk, hogy a bemenő információ *valamilyen mértékben* megjelenjen a kimenetekben. Ehhez a szerzők azt vizsgálták, hogy a bemeneti vektorok *párjai*, valamint a hozzájuk tartozó kimeneti értékek mennyire korrelálhatnak. A súlyok bizonyos eloszlásai a „rend fázisába” tartoznak, melyekre igaz, hogy minden bemeneti párhoz tartozó kimenetek aszimptotikusan korreláltak, és így ezek eltűnő gradienshez vezethetnek. Ezzel szemben más súlyeloszlások a „kaotikus fázisba” sorolódnak (ahol a megfelelő kimenetek aszimptotikusan dekorreláltak, és fölrobbanó gradienshez vezethetnek) (Poole és mtsai, 2016). A két eloszláshalmazt elválasztó határ a „káosz határa”, és kívánatos a kezdősúlyainkat egy ilyen eloszlás szerint választanunk (Schoenholz és mtsai, 2017).

A föntieket vitte tovább Hayou és munkatársai cikke (Hayou és mtsai, 2019), melyben elsősorban a különlegesen mély hálókra (10-200 rejtett réteg) koncentráltak. Megmutatták, hogy a korábbi inicializálási eljárások ilyen mélységben már nem vezetnek konvergenciához. Javaslatuk az volt, hogy a súlyok eloszlását

úgy kell megválasztani, hogy azok a „káosz határára" essenek, ami azt is jelenti, hogy (továbbra is 0 átlagon tartva azokat) a szórás értékét minden aktivációs függvényhez egyedileg (valamint a biasok szórásához igazodva) kell meghatározni. Glorot és He módszereivel összhangban a kapott értékeket továbbra is el kell osztani az adott réteg bemeneteinek számának négyzetgyökével. Kísérleti eredményeik alapján ez az eljárás lehetővé tette akár 200 rejtett réteget tartalmazó háló tanítását is tanh és ReLU aktivációs függvényekkel (Hayou és mtsai, 2019), ugyanakkor a módszer könnyűszerrel alkalmazható más függvényekre is.

3. A kísérletek technikai paraméterei

3.1. A tesztelt aktivációs függvények

Kísérleteinkben három aktivációs függvényt alkalmaztunk. Az első a hagyományos **szigmoid** függvény volt, melynek képlete

$$sig(x) = \frac{1}{1 + e^x}. \tag{5}$$

A következő aktivációs függvény, mely szintén igen elterjedt mind a beszédtechnológia, mind általában a gépi tanulás területén, a rectifier (vagy **ReLU**) függvény:

$$ReLU(x) = \max(x, 0). \tag{6}$$

Végül teszteltük a **szoftplusz** aktivációs függvényt is, melyet szokás a ReLU függvény folytonosan deriválható közelítésének is tartani (Dugas és mtsai, 2001):

$$softplus(x) = \log(1 + e^x). \tag{7}$$

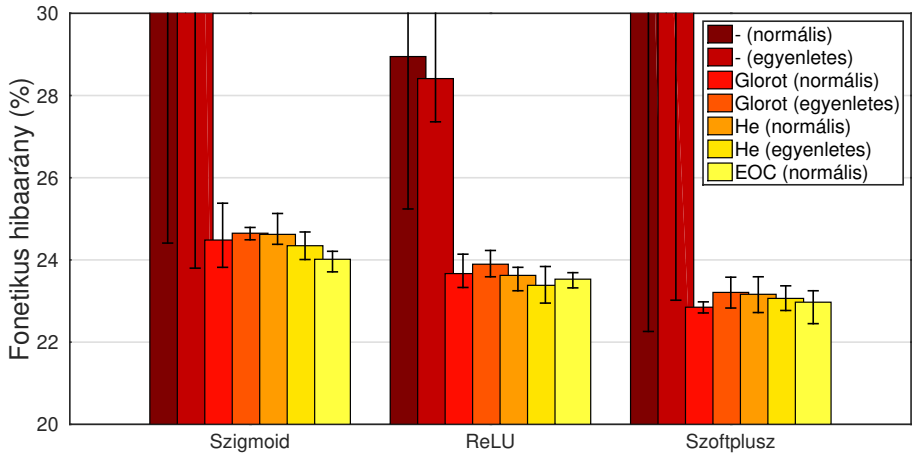
A kimeneti rétegben minden esetben a szoftmax függvényt alkalmaztuk.

3.2. A TIMIT adatbázis

Kísérleteinket az angol nyelvű TIMIT beszédadatbázison végeztük (Lamel és mtsai, 1986), mely relatíve kis mérete (kb. 3 óra) ellenére még mindig gyakran használt. A súlyokat a 3696 felvételből álló tanítóhalmaz közelítőleg 90%-án (3342 felvételen) tanítottuk, a fennmaradó 354 felvétel pedig a tanítási ráta vezérlésében kapott szerepet (*learn rate scheduling*). Mivel nem volt hangolandó hiperparaméterünk, a kiértékelést közvetlenül a 192 felvételből álló „mag" (core) teszthalmazon végeztük. Kiértékelés előtt a fonémacímkéket a bevett gyakorlatnak megfelelően 39 kategóriába vontuk össze (Lee és Hon, 1989).

3.3. DNN-paraméterek

Akusztikus neurális hálóink 5 rejtett réteget tartalmaztak, minden rejtett rétegben 1024 neuronnal. Bemenetként keretszinten egy 40 sávos mel-szűrőkészlet



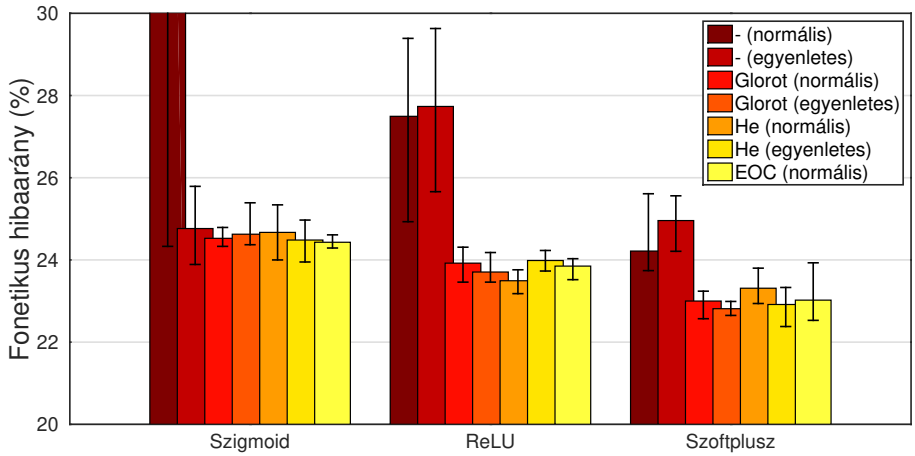
1. ábra: A különböző súlyinicializáló megközelítésekkel elért átlagos fonetikai hibaarányok **regularizáció nélkül** a TIMIT adatbázis „mag" teszthalmazán.

energiakimeneteit használtuk, a szokásos első és második derivált értékeivel kiegészítve; minden kerethez felhasználtuk a mindkét oldali szomszédos 8-8 keret jellemzővektorait is, így a hálók 2091 bemeneti neuronnal rendelkeztek. Kimenetként 858 kontextusfüggő kapcsolt állapotot használtunk, ennek megfelelő számú kimeneti neuronnal. A keresési lépést a HTK programcsomag (Young és mtsai, 2006) egy módosított változatával végeztük el.

Mivel kíváncsiak voltunk arra is, hogy az egyes inicializálási eljárások mennyire igénylik a tanítás során valamilyen regularizációs technika alkalmazását, minden kísérletünket megismételtük L2 regularizációval is. Mivel kísérleteink tárgya alapvetően a *véletlen* súlyinicializálási stratégiák hatékonysága volt, releváns volt a kapott eredmények stabilitása is, ezért minden konfigurációra öt különböző modellt tanítottunk (eltérő random seedek használatával).

4. Eredmények

A **regularizáció nélkül** elért átlagos fonémaszintű hibaarányok az 1. ábrán láthatók; a képen feltüntettük az öt tanított modell közül a legjobb és a legrosszabb eredményét is. Látható, hogy amennyiben standard súlyinicializálást használunk, az eredmények elég rosszak: az öt tanított modellből a szigmoid aktivációs függvényt alkalmazva csak két-két (normális és egyenletes eloszlású kezdősúlyok), míg a szoftplusz függvény esetén csak három és két modell tanult egyáltalán, így adódott az átlagos hiba ilyen magasnak. A ReLU függvény esetében ennél kedvezőbb volt a helyzet, de kompetitívnek ekkor sem tekinthetjük: normális eloszlású kezdősúyok esetén négy modell hibája $28,8 - 30,7\%$ közé esett, és csak egy esetben kaptunk elfogadható teljesítményt ($25,2\%$-os fonéma-hibaarány), míg egyenletes eloszlású súlyoknál mind az öt modell $27,3\%$ és $31,2\%$ közé eső fonetikai hibaarányokhoz vezetett.



2. ábra: A különböző súlyinicializáló megközelítésekkel elért átlagos fonetikai hibaarányok **L2 regularizációval** a TIMIT adatbázis „mag” teszthalmazán.

A fönnmaradó három tesztelt súlyinicializálási eljárás (Glorot, He és EOC) esetében azt látjuk, hogy nem szükséges a súlyok L2 regularizációja ahhoz, hogy használható fonémafelismerési teljesítményt kapjunk: minden esetben $23-24\%$ körüli átlagos fonetikai hibaarányokat tapasztaltunk. Az Edge of Chaos eljárás ugyan stabilan a legjobb két modell között volt, de a különbség nyilvánvalóan nem szignifikáns. A módszer előnye lehet ugyanakkor, hogy a szigmoid és a ReLU függvények esetében az öt tanított modell egymáshoz nagyon hasonló teljesítményhez vezetett. Ez azonban igen korlátozott előnynek tűnik, egyrészt mert ez pont az egyébként legalacsonyabb hibarátákhoz vezető szoftplusz aktivációs függvény esetében nem teljesült, másrészt mert a tanított modellek teljesítménye közötti különbség nagyobb tanítóadatbázis használata esetén eltűnhet.

Különbségek inkább az egyes aktivációs függvények esetében adódtak: látható, hogy a szigmoid függvények helyett érdemes a ReLU, de még inkább a szoftplusz függvényt alkalmazni. Természetesen az, hogy L2 vagy más regularizáció (pl. dropout) használata nélkül is lehetségesnek bizonyult egy öt rejtett rétegből álló neurális háló tanítása, már önmagában is érdekes tapasztalat (habár a javasolt súlyinicializáló eljárások motivációja éppen ez volt).

Az **L2 regularizáció használatával** elért átlagos fonémaszintű hibaarányokat a 2. ábrán tüntettük föl (ismét a legjobb és legrosszabb modellek teljesítményével együtt). Látható, hogy a regularizáció használata lehetővé tette a standard súlyinicializáló eljárás használatát a szoftplusz aktivációs függvény esetében is; a másik két függvény esetén azonban a helyzet nem változott (azaz a szigmoidnál még mindig használhatatlan, a ReLU esetében pedig még mindig egyszerűen csak rossz eredményeket kaptunk). Más tekintetben nagyon hasonlóak az eredmények a súlyregularizáció nélkül elértekhez. Véleményünk szerint ez azt jelenti, hogy (legalábbis a DNN akusztikus modellek esetén megszokott méretű hálók esetén) a súlyok megfelelően megválasztott kezdőértékei mellett a

gyakorlatban nincs szükség a tanítás során további regularizációra sem a vanishing gradient, sem az exploding gradient effektus elkerüléséhez. Természetesen a későbbiekben tervezzük ezt a következtetésünket mind nagyobb adatbázisokon, mind mélyebb hálók használata esetén ellenőrizni.

Inicializálás módja		**Szigmoid**		**ReLU**		**Szoftplusz**	
		—	L2	—	L2	—	L2
—	Normális	67.0%	66.6%	28.9%	27.5%	52.3%	24.2%
	Egyenletes	66.7%	24.8%	28.4%	27.7%	37.9%	25.0%
Glorot	Normális	24.5%	24.5%	23.7%	23.9%	22.8%	23.0%
	Egyenletes	24.7%	24.6%	23.9%	23.7%	23.2%	22.8%
He	Normális	24.6%	24.7%	23.6%	23.5%	23.2%	23.3%
	Egyenletes	24.3%	24.5%	23.4%	24.0%	23.1%	22.9%
EOC	Normális	24.0%	24.4%	23.5%	23.8%	23.0%	23.0%

1. táblázat. *A különböző megközelítések által elért átlagos fonetikai hibaarányok a TIMIT adatbázis „mag” teszthalmazán.*

Az átlagos fonetikai hibaarányokat az 1. táblázatba is kigyűjtöttük. Látható, hogy (a viszonyítási alapként szolgáló inicializálástól eltekintve) a hibaértékeket elsősorban az aktivációs függvény határozza meg: szigmoid esetén a $24,0-24,7\%$, ReLU esetén a $23,4-24,0\%$, szoftplusz esetén pedig a $22,8-23,3\%$ intervallumba estek.

5. Összegzés

Jelen tanulmányunkban különböző kezdősúly-meghatározási stratégiákat hasonlítottunk össze mély neurális hálós akusztikus modellek esetében. Vizsgálatainkban három különböző eljárással határoztuk meg a súlyok véletlen (normális, illetve egyenletes) eloszlásának szórását, míg annak átlagát minden esetben nullára állítottuk. Tesztjeink során három különböző aktivációs függvényt is megvizsgáltunk. Az előálló fonetikai hibaarányok alapján úgy véljük, érdemes valamelyik megvizsgált stratégiát alkalmazni a tanítás előtt, ugyanakkor az egyes eljárások pontossága között nem találtunk markáns különbségeket.

Tanulmányunkban azt is vizsáltuk, hogy a súlyok L2 regularizációja milyen hatással van a tanított DNN-ek teljesítményére. Tapasztalataink szerint amennyiben akár a Glorot, akár a He, akár az Edge of Chaos inicializálást alkalmazzuk, a súlyok tanítás közbeni regularizációja elhagyható. Ugyanakkor fontosnak érezzük megjegyezni, hogy kísérleteink egy kisméretű beszédadatbázison (a TIMIT-en) történtek; reálisnak tartjuk, hogy több tanító adat használatával az azonos paraméterekkel, csak eltérő random seeddel tanított modellek teljesítménye még jobban közelítsen egymáshoz. Egy másik érdekes lehetséges kutatási iránynak tartjuk a jelenleg általánosan használtnál lényegesen mélyebb (10-20,

akár 50-100) rejtett rétegű DNN akusztikus modellek tanítását, ez azonban szintén további vizsgálatokat igényel.

Köszönetnyilvánítás

Jelen kutatás eredményei az „Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein" című, EFOP-3.6.3-VEKOP-16-2017-0002 számú projekt támogatásával készültek. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (FK 124413), részben pedig az Innovációs és Technológiai Minisztérium (ITM 2018-1.2.1-NKP-2018-00008 és TUDFO/47138-1/2019-ITM) is támogatta. Gosztolya Gábor és Tóth László kutatásait az MTA Bolyai János Kutatási ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosítók: ÚNKP-19-4-SZTE-51) támogatta.

Hivatkozások

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., Garcia, R.: Incorporating second-order functional knowledge for better option pricing. In: Advances in Neural Information Processing Systems (2001)

Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Machine Learning Research. pp. 249–256 (2010)

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: International Conference on Machine Learning. pp. 369–376. Pittsburgh, PA, USA (2006)

Grósz, T., Gosztolya, G., Tóth, L.: Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler-divergencia alapú klaszterezéssel (in Hungarian). In: MSZNY. pp. 174–181. Szeged (2015)

Grósz, T., Gosztolya, G., Tóth, L.: Mély neuronhálós beszédfelismerők gmm-mentes tanítása (in Hungarian). In: MSZNY. pp. 170–180. Szeged (2017)

Hanin, B., Rolnick, D.: How to start training: The effect of initialization and architecture. In: Neural Information Processing Systems. Montréal, Kanada (2018)

Hayou, S., Doucet, A., Rousseau, J.: On the impact of the activation function on deep neural networks training. In: International Conference on Machine Learning (2019)

He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: IEEE International Conference on Computer Vision. pp. 1026–1034 (2015)

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer, S.C., Kolen, J.F. (szerk.) A Field Guide to Dynamical Recurrent Neural Networks (2001)

Lamel, L., Kassel, R., Seneff, S.: Speech database development: Design and analysisof the acoustic-phonetic corpus. In: DARPA Speech Recognition Workshop. pp. 100–109 (1986)

Lee, K., Hon, H.: Speaker-independent phone recognition using Hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing 37(11), 1641–1648 (1989)

Pennington, J., Schoenholz, S.S., Ganguli, S.: Resurrecting the sigmoid in deep learning throughdynamical isometry: Theory and practice. In: Neural Information Processing Systems. Long Beach, CA, USA (2017)

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., Ganguli, S.: Exponential expressivity in deepneural networks through transient chaos. In: Advances in Neural Information Processing Systems. pp. 3360–3368 (2016)

Pretorius, A., Van Biljon, E., Kroon, S., Kamper, H.: Critical initialisation for deep signal propagation in noisy rectifier neural networks. In: Neural Information Processing Systems. Montréal, Kanada (2018)

Schoenholz, S.S., Gilmer, J., Ganguli, S., Sohl-Dickstein, J.: Deep information propagation. In: International Conference on Learning Representations (2017)

Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)

Zhu, L., Kilgour, K., Stüker, S., Waibel, A.: Gaussian free cluster tree construction using Deep Neural Network. In: Interspeech. pp. 3254–3258. Drezda, Németország (Sep 2015)

A történetszerkezet automatikus elemzése és összefüggése az elbeszélő személy érzelmi intelligenciájával

Pólya Tibor

Eötvös Loránd Kutatóhálózat, Természettudományi Kutatóközpont
1117 Budapest, Magyar tudósok körútja 2.
polya.tibor@ttk.hu
Károli Gáspár Református Egyetem, Pszichológiai Intézet
1037 Budapest, Bécsi út 324.
polya.tibor@kre.hu

Kivonat: Az utóbbi 15 évben egyre gyakrabban alkalmazzák a természetesnyelv-feldolgozási eszközöket a történetek szerkezetének automatikus elemzésére. A pszichológia területén 2 elemzőrendszert is kidolgoztak a történetek elemzésére. A tanulmány 4 új eljárást mutat be, amelyek célja a történetszerkezet komplexebb elemzése. A 4 eljárás a történetek érzelmi jelentésének integrációját, a narratív transzformációkat, történetnyelvtani kategóriákat és a narratív értékelések előfordulását azonosítja. Az elemzési eljárások megbízhatósága alkalmassá teszi ezen eljárásokat arra, hogy pszichológiai szövegelemzésben felhasználjuk. Az elemzési eljárások validitását a történet konstrukció és a képességként meghatározott érzelmi intelligencia közötti kapcsolatot feltáró vizsgálat eredményei igazolják.

1 Bevezetés

Az elmúlt másfél évtizedben egyre gyakrabban alkalmazzák a természetesnyelv-feldolgozás területén kidolgozott elemzési eljárásokat a történetek elemzésében (összefoglalóan lásd például Mani, 2013). Az alkalmazások fókuszában a sajátosan a történetekre megjelenő nyelvi jelenségek állnak, így a történetek idői szerkezete (például Ehmann, 2004; Zhao és mtsai, 2012) és azon események azonosítása, amelyek a történet eseményvázát adják (például Elsner, 2012).

1.1 Történetelemző rendszerek a pszichológiai kutatásban: Narcoder és NarrCat

A történet specifikus jellemzőinek azonosítására kidolgozott elemzési eljárások egyik fontos alkalmazási területét a pszichológiai vizsgálatok adják. A pszichológiai vizsgálatok érdeklődését két tényező is magyarázza. Egyrészt a narratív pszichológiai megközelítés megjelenésével és elterjedésével egyre több olyan pszichológiai vizsgálat

készült, amelyek abból indulnak ki, hogy a mentális működés alapvető szerveződési elve a történetszerű szerveződés (Bruner, 1986; László, 2005; Pólya, 2007; Sarbin, 1986). Másrészt a narratív elemzési módszerek a narratív pszichológiai megközelítéstől függetlenül is hatékony eszköznek bizonyultak a személyes élmények empirikus vizsgálatában (például Angus és mtsai, 1999).

A történetelemzés elméleti és gyakorlati fontossága ellenére a pszichológia területén csak két olyan automatikus elemző eljárást dolgoztak ki ez idáig, amelyek a természetesnyelv-feldolgozás eszközeit használják a történetek sajátos szerveződési elveinek vizsgálatára. Az egyik eljárás a Narcoder, amelyet Stein és munkatársai (Stein és Hernandez, 2007) dolgoztak ki angol nyelvű történetek elemezésére. A másik eljárás a NarrCat, amelyet László János és munkatársai (Ehmann és mtsai, 2014; László és mtsai, 2013) dolgoztak ki magyar nyelvű történetek elemzésére. A két rendszer között számos eltérés van. A Narcoder a történetnyelvtani megközelítésre épít, amely a generatív grammatika elveit alkalmazva próbálja leírni a történetek szerveződését. A NarrCat ezzel szemben a történetstruktúra kompozicionális megközelítéséből indul ki és elsősorban a tartalomelemzés módszertanába illeszkedik. Annak ellenére, hogy a két eljárás a történetszerkezet részben azonos tartományait vizsgálja az elemzett szerkezeti jellemzők listájában is vannak eltérések (lásd 1. Táblázat). Mindkét eljárás elemzi a szereplők jellemzőit, cselekvéseit és belső tudattartalmaikat is. Azonban míg a Narcoder a szereplők személyes diszpozícióit és preferenciáit azonosítja, a NarrCat a szereplők két kategóriáját azonosítja a szövegben, az önéletrajzi elbeszélésekben az elbeszélővel azonos szereplő említését, a csoporttörténetekben pedig a releváns csoportra való utalást. A cselekvések területén a Narcoder az események közötti cél alapú szerveződést ragadja meg a Kezdő esemény, Cél, Cselekvési terv, Cselekvés és Kimenet kategóriáinak azonosításával. Ezzel szemben a NarrCat a szereplők cselekvéseinek ágencia szintjét elemzi az Aktív és a Passzív cselekvések, az Intenciók és a cselekvésre vonatkozó Korlátozások azonosítása révén. Az aktív cselekvést a szereplő szándékosan hajtja végre, míg a passzív cselekvés hatásait elszenvedi a szereplő. A belső tudattartalmak területén a Narcoder nagyon differenciáltan azonosítja a szereplők mentális állapotát, a NarrCat ezzel szemben szereplők kognitív és érzelmi állapotait azonosítja. Mindezek mellett a NarrCat további narratív jellemzőket is vizsgál, így az értékelést, a tagadást és a történet téridői perspektíváját. Az utóbbi jellemző meghatározásának alapját a tartalomelemzési módszer helyett a történetszerkezet strukturalista leírásai (például Bal, 1997) adják.

A számos eltérés ellenére az elemzésből nyert adatok feldolgozásában a két eljárás megegyezően jár el. Ugyanis mindkét eljárás az egyes kategóriák előfordulási gyakoriságával jellemzi a történeteket és nem vizsgálják azt, hogy a narratív kategóriák hogyan kapcsolódnak egymással. Ezzel az elemzési stratégiával mindkét eljárás jelentősen korlátozza a narratív szerkezetnek azt a tartományát, amelyet képesek elemezni. A történetek szerkezetének ugyanis az is fontos összetevője, hogy a narratív kategóriák hogyan kapcsolódnak egymáshoz. Ebben a cikkben olyan narratív elemzési eljárások fejlesztését mutatjuk be, amelyek célja, hogy a narratív kategóriák közötti kapcsolatok azonosítása révén elemezzék a történetek szerveződését és ezáltal tágítsák a narratív szerveződésnek azt a tartományát, amelynek elemzésében felhasználhatók automatikus elemzési eljárások.

1. táblázat: A Narcoder és a NarrCat által elemzett narratív kategóriák

Narratív kategóriák	Narcoder	NarrCat
Szereplő	Személyes diszpozíció	Személyreferencia
	Preferencia	Egyes szám első személyű
		Többes szám első személyű
Cselekvés	Kezdő esemény	Ágencia
	Cél	Aktív cselekvés
	Cselekvési tervek	Passzív cselekvés
	Cselekvés	Szándék
	Kimenet	Korlátozás
Belső tudattartalmak	Érzelmi állapot	Pszichológiai perspektíva
	Általános affektív állapot	Kognitív állapot
	Hangulati állapot	Érzelem
	Mentális állapot	
	Vélekedés	
	Vélekedés veszélyeztetése	
	Vélekedés átalakítása	
Értékelés		Értékelés
Tér-idői perspektíva		Tér-idői perspektíva
Tagadás		Tagadás

2 Narratív elemző eljárások

2.1 Az érzelmi jelentés integrációja

Az érzelmi jelentés integrációját elemző eljárás fejlesztésének célja az, hogy feltárjuk az érzelmi konnotációval rendelkező szavak közötti kapcsolat minőségét. Az érzelmi konnotációban pozitív és negatív érzelmi jelentést különböztetünk meg, az érzelmi konnotációval bíró szavak közötti kapcsolat pedig azonos vagy ellentétes lehet. Ezen kategóriák felhasználásával az érzelmi jelentés integrációját az jelzi, ha két pozitív vagy két negatív érzelmi konnotációjú szó azonosságot kifejező kötőszóval van összekapcsolva, vagy egy pozitív és egy negatív érzelmi konnotációjú szó ellentétes kapcsolatot kifejező kötőszóval van összekapcsolva. Ezzel szemben az érzelmi jelentés diszintegrációját jelzi, ha két pozitív vagy két negatív érzelmi konnotációjú szó ellentétes kapcsolatot kifejező kötőszóval van összekapcsolva, vagy egy pozitív és egy negatív érzelmi konnotációjú szó azonos kapcsolatot kifejező kötőszóval van összekapcsolva.

Az érzelmi jelentés integrációja elemző eljárás kidolgozásához a magyar nyelv leggyakrabban használt 4400 igéjének, főnevének és melléknevének szótövét kódolta 3 független kódoló 7 fokú skálán. A skála értékekből átlagot számoltunk és a hármas értéknél alacsonyabb átlagú szavak kerültek a negatív érzelmi konnotációjú szavak csoportjába (494 szótő), míg az ötös értéknél magasabb átlagú szavak kerültek a pozi-

tív konnotációjú szavak csoportjába (1053 szó). A leggyakrabban használt 68 kötőszót szintén 3 független kódoló sorolta 3 csoportba: az azonos (23 kötőszó, például *és, meg*), az ellentétes (28 kötőszó, például *de, pedig*) kapcsolatot kifejező kötőszavak csoportjába, illetve azon kötőszavak csoportjába ahol az azonosság vagy ellentétesség nem értelmezhető (17 kötőszó, például *amikor, azért*).

2.2 Narratív transzformációk

A narratív transzformációk elemzésének kiindulópontját Bruner (1986) elképzelése adta, amely a történetet két tartomány együtteseként írja le. Bruner szerint az egyik tartományt a történetbe foglalt cselekvések adják, a másik tartományt pedig a szereplők belső tudattartalmai. A történetekben jellemzően mindkét tartomány jelen van, így a cselekvések leírása két tartományban is megjelenhet. Az, hogy egy cselekvés a cselekvések tartományában vagy a tudattartalmak tartományában jelenik meg a narratív transzformációk jelenlététől függ, mivel a narratív transzformációk képesek arra, hogy a cselekvés leírását a cselekvések tartományából átemelje a tudattartalmak tartományába. A narratív transzformációk első szisztematikus leírását Todorov (1977) készítette el. Todorov 12 narratív transzformációt írt le, amelyeket két csoportba, az egyszerű és az összetett transzformációk csoportjába sorolt (lásd 2. Táblázat). Az egyszerű transzformációk esetében egy szó, míg az összetett transzformációk esetében egy tagmondat az, amely átemeli a cselekvést a tudattartalmak tartományába.

2. táblázat: Narratív transzformációk Todorov (1977) alapján

Transzformációk	**Meghatározás**
Egyszerű transzformáció	
Modalitás	A cselekvés lehetőségének vagy szükségességének bemutatása
Intenció	Szándék a cselekvés végrehajtására
Eredmény	A cselekvés végrehajtásának hangsúlyozása
Mód	Cselekvés kivitelezésének módja
Aspektus	A cselekvés idői lefutásának jelzése
Státusz	cselekvés tagadása
Összetett transzformáció	
Megjelenés	Cselekvés végrehajtásának látszata
Tudás	Cselekvés tudatosulása
Leírás	Beszédaktusok
Feltevés	Cselekvés végrehajtása a jövőben
Személyhez kapcsolás	Cselekvés beágyazása a szereplő tudattartalmába
Attitűd	Cselekvés értékelése

A narratív transzformációk elemző eljárás kidolgozása során összegyűjtöttük azokat a szavakat és kifejezéseket, amelyek leggyakrabban részt vesznek a cselekvés tudattartalmak tartományába való átemelésében. Ehhez figyelembe vettük a szavak jelentését, szófaját, morfológiai jellemzőiket és szintaktikai kapcsolataikat.

2.3 Történetnyelvtan

A történetnyelvtanok a generatív grammatikai elveit alkalmazva újraírószabályok felhasználásával írja le a történetek szerkezetét. A történetekre való alkalmazás esetében azonban nem a mondatok szerkezetét írjuk le a mondat szavai közötti kapcsolat tisztázásával, hanem a történetek szerkezetét írjuk le a történetbe foglalt cselekvések és állapotleírások közötti kapcsolatok tisztázásával. A 3. táblázatban például az a 11 újraírószabály látható, amelyek Rumelhart (1975) használt fel a történetek szerkezetének leírásához.

3. táblázat: Rumelhart (1975) történetnyelvtanának újraírószabályai

1. Történet	→	Helyzetleírás + Epizód
2. Helyzetleírás	→	Állapot*
3. Epizód	→	Esemény + Reakció
4. Esemény	→	Állapotváltozás / Cselekvés / Esemény
5. Reakció	→	Belső válasz + Nyílt válasz
6. Belső válasz	→	Érzelem / Vágy
7. Nyílt válasz	→	Cselekedet / Kísérlet
8. Kísérlet	→	Terv + Megvalósítás
9. Megvalósítás	→	Cselekvés + Következmény
10. Cselekedet	→	Alcél + Kísérlet*
11. Következmény	→	Reakció / Esemény

A történetnyelvtani elemzést végrehajtó elemző eljárás kidolgozásának első lépéseként a 2. szabályban szereplő Helyzetleírás és a 9. szabály által megfogalmazott Cselekvés és Következmény kategória automatikus felismerését valósítottuk meg. Ehhez ebben az esetben is a szavak jelentését, szófaját, morfológiai jellemzőiket és szintaktikai kapcsolataikat vettük figyelembe.

2.4 Narratív értékelés

A narratív értékelések elemzésének alapját Labov és Waletzky (1967) modellje adta, amely a diszkurzív kontextusba illeszkedő történetek szerkezetét írja le. A modell egyik makroszerkezeti kategóriája az értékelés. Labov (1972) későbbi munkájában részletes leírást adott a narratív értékelés típusairól és a narratív értékelést megvalósító nyelvi eszközökről (lásd 4. táblázat).

A narratív értékelést elemző eljárás kidolgozásához a Kvantifikáció, Kérdés, Felszólítás, Explicit összehasonlítás, Magyarázó tagmondat hozzáadása és Minősítés azonosítására dolgoztunk ki eljárásokat. Ebben az esetben a szavak jelentését, szófaját, morfológiai jellemzőiket, szintaktikai kapcsolataikat és a központozást vettük figyelembe.

4. táblázat: A narratív értékelést megvalósító nyelvi elemek Labov (1972) alapján

Narratív értékelés típusa	Nyelvi elemek
Intenzifikálás	Gesztikuláció
	Kifejező fonológia
	Kvantifikálók
	Cselekvés ismétlése
	Jellemző szófordulat
Összehasonlítás meg nem történt eseményekkel	Cselekvés tagadása
	Jövőbeli események
	Kérdés
	Felszólítás
	Explicit összehasonlítás
Összekapcsolás megtörtént eseménnyel	Folyamatos ige
	Több összekapcsolt folyamatos ige
	Két leíró főnév
	Két melléknév
Magyarázó tagmondat hozzáadása	Minősítés
	Cselekvés okának leírása
	Egyéb hozzáadás
	Eredmény: az eseménysort lezáró esemény

3 A narratív elemző eljárások implementációja és bemérése

A 4 narratív elemző eljárás implementációja során a következő elemzési lépéseket kapcsoltuk össze. A szöveg nyelvi elemzését a magyarlanc (Zsibrita és mtsai, 2013) programcsomag végzi el. Az elemzést megvalósító kereső algoritmusokat a NooJ számítógépes fejlesztési környezetben (Silberztein, 2004) készítettük el. A történetelemzés eredményeinek statisztikai elemzését pedig a WordStat tartalomelemző programmal végeztük el.

Az elemző eljárások kiértékelését 20 önéletrajzi történet kézi elemzésével összehasonlítva végeztük el. Az elemzett történetek teljes szószáma 2 700 szó volt. A történeteket 3, a narratív kategóriák alkalmazásban tapasztatot szerzett kódoló elemezte egymástól függetlenül. Az egyetértés mértékének megállapításához a Cohen-féle kappa értéket számoltuk ki (0,85). A kódolások közötti eltéréseket közös megegyezéssel oldották fel a kódolók. Az így kapott kódolást tekintettük a történetelemzés gold standardjának.

A gépi kódolás megbízhatóságát a találat és a pontosság mérőszámaival jellemezzük. A találat és pontosság mutatója az Érzelmi jelentés integrációja esetén 76 és 82, a narratív transzformációk esetén 84 és 81, a két történetnyelvtani kategória esetén 79 és 82, végül a narratív értékelés esetében 76 és 79 % volt. A pszichológiai szövegelem-

zésekben konszenzuálisan a 80 % fölötti értéket fogadják el annak kritériumaként, hogy az elemzési eljárás megbízhatósága elfogadható (Gottschalk és Bechtel, 2008). Mivel az általunk kidolgozott narratív elemző eljárások megbízhatósági mutatói meghaladják, illetve megközelítik ezt az értéket, az elemzési eljárások megbízhatóan alkalmazhatók a pszichológiai vizsgálatokban.

4 Pszichológiai vizsgálat

A narratív elemzési eljárások pszichológiai validitásának vizsgálatához az érzelmi intelligencia konstruktumát használtuk. Az érzelmi intelligencia az érzelmi vonatkozású információk feldolgozásában való jártasságot jelenti (MacCann és Roberts, 2008). A pszichológiai kutatásokban kétféle meghatározása van a fogalomnak. Az egyik képességként, a másik személyiségjellemzőként határozza meg az érzelmi intelligenciát. A két meghatározás az érzelmi intelligencia mérési módjában is eltér. A személyiségjellemzőként meghatározott érzelmi intelligencia kérdőívekkel mérhető, míg a képességként meghatározott érzelmi intelligencia méréséhez olyan feladatra van szükség, amelynek megoldásához szükség van erre a képességre. Álláspontunk szerint az érzelmi epizódokról beszámoló történetek konstrukciójához szükséges az érzelmi intelligencia képessége. Ennek alapja az, hogy a történetek rendszerint érzelmi vonatkozású információt is tartalmaznak, ezért a történetek konstrukciójában megnyilvánul az érzelmi intelligencia képessége. Ez alapján azt várjuk, hogy a képességként meghatározott érzelmi intelligencia szintje kapcsolatban lesz a történetek szerkezetét elemző eljárások eredményével.

A vizsgálatban 90 fő vett részt. Minden résztvevő 4 érzelmi epizódot mesélt el: egy emlékezetes bulit, egy stresszes vizsgát, egy baráttal való megismerkedést és egy személyes konfliktust. A résztvevők érzelmi intelligenciájának érzelemszabályozás komponensét az Érzelemszabályozás Szituációs Teszt (MacCann és Roberts, 2008) magyar változatával (Nagy és mtsai, 2015) mértük fel. A teszt 44 tételből áll, minden tétel egy szituációt ír le, és a kitöltőnek négy válaszlehetőség közül kell kiválasztani az adott helyzetben leghatékonyabb reakciót. Az érzelmi intelligencia differenciált mérése érdekében szintén kérdőíves eljárásokkal felmértük a résztvevők klasszikus intelligenciáját (Raven 1981; Raven és Rózsa 2006), verbális képességét (Smith és Whetton 1988; Rózsa, 2007) és személyiségjellemzőit (Carpara és mtsai, 1993; Rózsa és mtsai, 2006) is.

A vizsgálatban 360 történetet elemeztünk, ezek teljes terjedelme 51 860 szó volt, egy történet átlagos hossza 144,1 (SD=81,2) szó volt. A történeteket a 4 narratív elemző eljárással elemeztük.

A vizsgálat eredményei számos összefüggést tártak az érzelmi intelligencia szintje és a történetek szerkezete között. Az érzelmi intelligencia szintje pozitívan korrelál az integrált érzelmi jelentés előfordulásának gyakoriságával ($r=.13$; $p < .01$), a narratív transzformációk közül az Attitűd ($r=.10$; $p < .05$) és az Aspektus ($r=.12$; $p < .01$) kifejezésének gyakoriságával, a történetnyelvtani kategóriák közül a Következmény előfordulási gyakoriságával ($r=-.14$; $p < .01$) és a narratív értékelések közül a Minősítés ($r=.10$; $p < .05$) előfordulási gyakorisága között.

Mindezek az eredmények alátámasztják a képességként meghatározott érzelmi intelligencia és a történetkonstrukció közötti kapcsolat meglétét, és így igazolják az itt bemutatott narratív elemzési eljárások pszichológiai validitását.

5 Összegzés

A tanulmányban 4 olyan automatikus elemzési eljárást mutattunk be, amelyek célja a komplex narratív szerkezet elemzése volt. A kifejlesztett szabályalapú elemzési eljárások képesek megbízhatóan megvalósítani ezt a feladatot. Vizsgálatunk eredményei a feltárt narratív szerkezet pszichológiai validitását is igazolják. Ugyanakkor az itt bemutatott elemzési eljárások önmagukban még nem valósítják meg a kutatás végső célját, de megteszik azt a fontos lépést, hogy egymással összekapcsolt kategóriákat elemeznek: az érzelmi jelentés integrációja modul a valenciával rendelkező kifejezések közötti kapcsolatot, a történetnyelvtani elemzés a célszerkezet megvalósulását, a narratív transzformációk modul a cselekvések tudattartalmakba ágyazását, végül a narratív magmondatok és narratív értékelés modul a cselekvések és értékeléseik közötti kapcsolatot. A narratívumok komplex struktúrájának feltárásához a jövőben klaszterelemzést, illetve a kategóriák előfordulásának idői mintázatát leíró elemzéseket tervezünk végezni.

Köszönetnyilvánítás

A Projekt a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatásával az NKFI Alapból valósult meg (K124206).

Hivatkozások

Angus, L., Lewitt, H., & Hardtke, K. The narrative process coding system: Research applications and implications for psychotherapy practice. *Journal of Clinical Psychology*, 55(10), 1255-1270. (1999)

Bal, M. *Narratology. Introduction to the theory of narrative*. 2nd edition. University of Toronto Press, Toronto (1997)

Bruner, J. S. *Actual minds, possible worlds*. Harvard University Press, Cambridge, (1986)

Carpara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. The „big five questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences, 15(3),* 281-288. (1993)

Ehmann B., Csertő I., Ferenczhalmy R., Fülöp É., Hargitai R., Kővágó P., Pólya T., Szalai K., Vincze O., László J. (2014). Narratív kategoriális tartalomelemzés: A NARRCAT. In: X. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY pp. 136–147. Szegedi Tudományegyetem, Szeged (2014)

Ehmann B. Elbeszélt élettörténeti epizódok időstruktúrája. In László J., Kállai J., & Bereczkei T. (Szerk.), *A reprezentáció szintjei* pp. 357–372. Gondolat Kiadó, Budapest (2004)

Gottschalk, L. A., & Bechtel, R. J (Eds.) *Computerized content analysis of speech and verbal texts and its many applications.* Nova Science Publisher, New York (2008)

Labov, W. & Waletzky, J. Narrative Analysis: Oral Versions of Personal Experience. In J. Helms (Ed*), Essays on the verbal and visual arts*. pp. 4-44. University of Washington Press, Seattle (1967)

Labov, W. *Language in the inner city*. pp. 354-396. Blackwell, Oxford (1972)

László, J., Csertő, I., Fülöp, É., Ferenczhalmy, R., Hargitai, R., Lendvai, P., Péley, B., Pólya, T., Szalai, K., Vincze, O., & Ehmann, B. Narrative Language as an Expression of Individual and Group Identity: The Narrative Categorical Content Analysis: *SAGE Open* April-June, 1-12. (2013)

László J. *A történetek tudománya. Bevezetés a nartatív pszichológiába.* Új Mandátum Kiadó, Budapest (2005)

Mani, I. *Computational modeling of narrative*. Synthesis Lectures on Human Language Technologies 18. Morgan & Claypool, Toronto (2013)

MacCann, C., & Roberts, R.D. New Paradigms for Assessing Emotional Intelligence: Theory and Data. *Emotion, 8(4),* 540–551. (2008)

Elsner, M. Character-based kernels for novelistic plot structure. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics Avignon, France, April 23-27, (2012)

Nagy H., Magyaródi T., & Sellei B. A képességalapú érzelmi intelligencia: új paradigmák a tesztfejlesztésben és pontozásban. Hazai tapasztalatok az érzelemmegértés és érzelemszabályozás szituációs tesztekkel. *Magyar Pszichológiai Szemle, 70(4/7).* 827–846. (2015)

Pólya T. Identitás az elbeszélésben. Szociális identitás és narratív perspektíva. Új Mandátum Kiadó, Budapest (2007)

Raven, J. *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1.* Harcourt Assessment, San Antonio (1981)

Raven, J., & Rózsa S. *Raven-féle progresszív mátrixok: Kézikönyv*. OS Hungary Tesztfejlesztő, Budapest (2006)

Rózsa, S, Kő, N, & Oláh, A. Rekonstruálható-e a Big Five magyar mintán: A Carpara-féle „Big Five Kérdőív" (BFQ) felnőtt változatának hazai adaptációja és nemzetközi összehasonlító elemzése. *Magyar Pszichológiai Szemle*, *26(1),* 57-76. (2006)

Rózsa S. *Általános Képesség Teszt*. OS Hungary Tesztfejlesztő. Budapest (2007)

Rumelhart, D. E. Notes on a schema for stories. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitivescience.* Academic Press. New York (1975)

Sarbin, T. R. *Narrative psychology. The storied nature of human conduct*. Praeger, New York (1986)

Silberztein, M. NooJ: A Cooperative, Object-Oriented Architecture for NLP. In: C. Muller, J. Royauté, M. Silberztein, (Eds.), *INTEX pour la Linguistique et Traitement Automatique des Langues.* pp. 359-370. Presses Universitaires de Franche-Comté, Besançon (2004)

Smith, P., & Whetton, C. *General abilities tests: User's guide.* NPER-Nelson, Windsor (1988)

Stein, N. L., & Hernandez, M. V. Assessing Understanding and Appraisals During Emotional Experience. In J. A. Coan, and J. J. B. Allen, (Eds), H*andbook of Emotion Elicitation and Assessment*. pp. 298—317. Oxford University Press, Oxford (2007)

Todorov, T. Narrative transformations. In *The Poetics of Prose*. Pp. 218-233. Cornell University Press. Ithaca (1977)

Zhao, R., Xuan, Q., & Roth, D. A robust shallow temporal reasoning system. NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstration Session pp. 29-32. (2012)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pp. 763–771. Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA (2013)

Kulcsfogalmak jelentésváltozása a Kádár-korszak politikai diskurzusában

Ring Orsolya[1], Kmetty Zoltán[1], Szabó Martina Katalin[1,2], Kiss László[1], Nagy Balázs[2], Vincze Veronika[3]

[1]Társadalomtudományi Kutatóközpont, CSS-RECENS Kutatócsoport
1097 Budapest, Tóth Kálmán u. 4.
[2]Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.
{kiss.laszlo, kmetty.zoltan, ring.orsolya}@tk.mta.hu,
{bnagy, martina, vinczev}@inf.u-szeged.hu

Kivonat A jelen dolgozatban a Magyar Szocialista Munkáspárt Központi Bizottságának (MSZMP KB) hivatalos havilapját, a *Pártélet* című kiadványt elemezzük néhány korabeli kulcsfogalom időbeli szemantikai változása szempontjából. A vizsgálatokhoz a korpuszt magunk hoztuk létre a lap teljes, digitalizált képként rendelkezésre álló anyagából. A korpusz egyedülálló, hiszen tudomásunk szerint nincs másik olyan digitalizált adatbázis, amely a Kádár-korszak longitudinális szövegbányászati elemzését lehetővé tenné. Mivel a *Pártélet* az állampárt hivatalos lapja volt, szövegeinek elemzése révén a korabeli politikai diskurzus megismerése, változásainak feltárása válik lehetővé. Jelen kutatásunk középpontjában a *döntés* és az *irányítás*, illetve a velük kapcsolatban álló fogalmak szemantikai tartalmának időbeli változása állt.
Kulcsszavak: korpuszépítés, információkinyerés, szóbeágyazási modellek, történeti diskurzuselemzés, számítógépes történettudomány

1. Bevezetés

A magyar történelem 1956 és 1989 közötti időszaka a történettudományban és a társadalomtudományban is a gyakran vizsgált korszakok közé tartozik. Politikai diskurzusának nyelvi jellemzői azonban eddig nem képezték elemzés tárgyát. E probléma okán korpuszt építettünk a *Pártélet* című folyóiratból, az MSZMP KB hivatalos ideológiai lapjából, majd azt NLP-módszerekkel feldolgoztuk és elemeztük.

A *Pártélet* című lap lehetővé teszi az országot irányító állampárt hivatalos diskurzusának elemzését. A kiadvány célja a politikai ideológia terjesztése, tehát a közvetlen agitáció és propaganda volt. A *Pártélet*, amely 54 150 példányban jelent meg, elsősorban nem az átlagemberekhez, hanem az állampárt különböző tisztségviselőihez szólt. A lap a párthierarchia egészét célozta, az első lapszámban megjelent ajánlás szerint ezen belül is elsősorban a pártfunkcionáriusok, az

aktivisták, valamint a propagandatevékenységért felelős pártmunkások számára íródott.

Az utóbbi években a magyarországi társadalmi és politikai folyamatokhoz kapcsolódó diskurzusok kvalitatív történelmi elemzése egyre gyakoribbá vált, ugyanakkor ezek az elemzések leginkább a hagyományos történeti diskurzuselemzés módszereit alkalmazzák, azaz a kortárs dokumentumok kvalitatív és manuális elemzésén alapulnak (Szabó, 2007; Pap, 2017; Gyáni, 2016). A szövegbányászati módszerek alkalmazását, illetve a kvantitatív elemzéseket alapvetően a digitalizált szövegkorpuszok hiánya akadályozta, ami pedig a digitális formátumú szövegek hiányával, valamint a történeti korpuszok építésének tipikus technikai problémáival mutat szoros összefüggést. Felismerve ezt a jelentős hiányt kezdtünk bele egy nagyméretű, digitalizált szövegkorpusz létrehozásába. A Pártéletkorpusz egyedülálló lehetőséget kínál számos, eleddig kivitelezhetetlen vizsgálat elvégzésére. Így például a segítségével elemezhető a korszak politikai diskurzusában zajló időbeli változások dinamikája (Xu és Kemp, 2015; Jatowt és Duh, 2014; Kulkarni és mtsai, 2014; Hamilton és mtsai, 2016a,b; Garg és mtsai, 2018).

Dolgozatunkban egy esettanulmányon, néhány kulcsfogalom időbeli dinamikájának a vizsgálatán keresztül mutatjuk be a korpuszunk és a kvantitatív szövegelemzés hasznosságát a történeti diskurzuselemzés számára. A munka során a szóbeágyazás módszerét alkalmazzuk, ami a természetesnyelv-feldolgozás (NLP) és a gépi tanulás területén gyakorta használt eszköz bármely két szó szemantikai kapcsolatának feltárására, illetve dinamikus perspektívába helyezve az időbeli szemantikai változások mérésére.

Esettanulmányunk célja a korszak különböző kulcsfogalmaival kapcsolatos politikai diskurzus változásainak azonosítása a Kádár-korszak éveiben Magyarországon. Ezen kulcsfogalmakat történelmi és szociológiai kritériumok alapján választjuk ki, és azt vizsgáljuk, hogy hogyan változik közöttük az időben a szemantikai kapcsolat. Mindehhez hat, történettudományi szempontból elkülönülő alkorszakot definiálunk, és az egyes korszakok vektorainak összehasonlításával kiszámítjuk a fogalmak időbeli dinamikáját.

Megvizsgálva a választott kulcsfogalmaknak a politikai diskurzusban betöltött szerepét, új, kvantitatív kutatási eredményekkel egészítjük ki és pontosítjuk a korábban e témában született kvalitatív eredményeket.

2. Történelmi háttér

Az 1956-os forradalom, majd az azt követő megtorlás időszaka után a Kádár-korszak a társadalom konszolidálását tűzte ki célul. A konszolidációs politika lényege a társadalom „lecsendesítése", a politikától, a politikai gondolkodástól való eltávolítása volt. A konszolidáció központi elemét képezte a fogyasztásra helyezett hangsúly, a társadalom széles rétegei számára elérhető „második gazdaságbeli" termelési formák tolerálása, idővel támogatása. Természetesen a sikeres konszolidációs politika az előző, Rákosi-korszakkal való viszonylagos szembehelyezkedést is szükségessé tette, akárcsak a hruscsovi Szovjetunió számára a sztálini előzményekkel való leszámolást. A már 1962-ben megindult új gazda-

ságpolitikai intézkedések előkészítették a terepet az 1968-ban kihirdetett gazdasági reformprogramnak, az „új gazdasági mechanizmusnak". Ennek keretében az egyes gazdasági szereplők, vállalatok a korábbinál lényegesen nagyobb önállóságra tehettek szert, döntési jogkörük és a központi irányítástól való függetlenségük megnőtt. Jelentősen megerősödött a „második szektor" (a saját fogyasztásra és értékesítésre termelő háztáji és kisegítő gazdaságok, a gazdasági munkaközösségek stb.), valamint a legális magánszektor is. A társadalom egyre jelentősebb rétegei érezhették úgy, hogy fogyasztási színvonaluk és életszínvonaluk javul. A reformfolyamat „keményvonalas" ellenzői azonban 1972-re megbuktatták a „mechanizmust", a gazdaságban ismét erőteljes központosítást indítottak. A döntések ismét centralizáltak lettek, a gazdasági szektor központi irányítása fokozódott. Az 1979-es „második olajárrobbanás" után újra bevezették az 1968. évi reform néhány elemét. Csökkent a központi irányítás szerepe és újra erősebben támogatták a lakosság „második gazdaságban" való részvételét. Mindez természetesen ismét csak a konszolidációs társadalompolitikával hozható összefüggésbe. A Rákosi-korszak kvázi háborús ideológiájával, háborús készülődésre utaló társadalompolitikájával szemben a Kádár-rendszer a „békés" szocialista fejlődést, a magas (illetve magasan tartott) fogyasztási színvonalat, valamint a depolitizálást tűzte zászlajára. A fogyasztás fokozása, a fogyasztási színvonal magasan tartása jelentette a Kádár-rendszer legfőbb erejét – egyfelől távol tudta tartani a társadalom jelentős rétegeit az aktív politizálástól (ez természetesen igen komoly ideológiai háttérmunkát igényelt), másfelől a rendszer „hatékonyságának" is bizonyítékául szolgált.

3. A korpusz létrehozása

3.1. Korpuszépítés, előfeldolgozás

A vizsgálatokhoz használt korpuszt, a Pártélet című lap számait az Arcanum Digitheca[1] oldalról töltöttük le. A lap szkennelt, PDF-formátumú oldalait a letöltés után további komplex feldolgozási folyamatoknak vetettük alá, amelyek eredményeképpen megkaptuk a szövegek elemezhető és megfelelő minőségű nyers változatait.

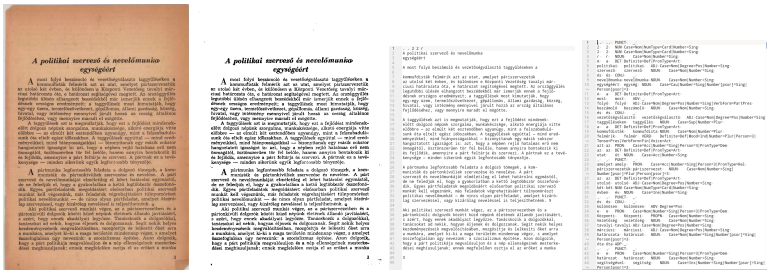
Először, mivel az optikai karakterfelismerő eszköz (Optical Character Recognition, OCR) képfájlokkal működik, az egyes PDF-oldalakat képi formátumba (PNG) konvertáltuk a pdftoppm konverter segítségével.

Második lépésként a PNG fájlokat binarizáltuk, vagyis fekete-fehér képekké alakítottuk át az ImageMagick[2] nevű eszközzel, amely a pdftoppm konverterhez hasonlóan ugyancsak minden Linux disztribúcióban elérhető. 50%-os küszöbértéket alkalmaztunk, ami azt jelenti, hogy minden ezen érték feletti pixelt feketére, a többit fehérre állítottuk. Ez a technika növeli az OCR-folyamat hatékonyságát azáltal, hogy növeli a kontrasztot a szöveg és a háttér között.

[1] https://adtplus.arcanum.hu/en/collection/Partelet/
[2] https://imagemagick.org

Az eredményfájlokon ezután a *tesseract* nevű nyílt forráskódú OCR eszközzel dolgoztunk tovább[3]. Az OCR segítségével az oldalfényképeket géppel olvasható szövegekké alakítottuk, alkalmassá téve őket a további gépi feldolgozásra. Végül a nyers szövegekből eltávolítottuk az oldalszámokat, az üres sorokat és kezeltük az elválasztásokat. Mindehhez saját bash és Python szkripteket használtunk.

A munkafolyamat egyes lépéseit az alább ábrák szemléltetik.



1. ábra: A fájlok állapota az egyes feldolgozói szakaszokban

A kapott szöveget a magyarlanc nyelvi elemző eszközzel[4] (Zsibrita és mtsai, 2013) dolgoztuk fel, amelynek segítségével a korpusz szövegeit először mondatokra bontottuk, tokenizáltuk és lemmatizáltuk. Ezután eltávolítottuk az írásjeleket és a stopszavakat. A stopszavak szűrése a további, szóbeágyazási modellel végzett szemantikai vizsgálatok szempontjából fontos lépés volt, hogy elkerüljük a nagyon gyakran előforduló szavak által okozott zajt az eredményekben.

A magyarlanc eszköz körülbelül 22000 szónak „ismeretlen szófaj" jelölést adott, azaz nem tudta meghatározni azok szófaját és morfológiai sajátságait, nagyrészt az OCR-hibáknak köszönhetően. Ezek kezelésére az alábbi lépéseket követtük. Először is kigyűjtöttük azokat az ismeretlen elemzésű szavakat, amelyek legalább hússzor előfordultak a korpuszban, továbbá minimum három karakterből álltak (pl. *imperializmus*). A következő lépésben ezeket kézi erővel javítottuk, majd az eredeti szövegben lecseréltük az eredeti alakokat a javított változatokra, végül újraelemeztük a korpuszt a magyarlanccal. Ezzel a módszerrel az ismeretlen szavak 64%-át sikerült kijavítanunk.

3.2. Alapvető korpuszadatok

A teljes *Pártélet* folyóirat összesen 33 évfolyamból áll, amelyeket 1956 és 1989 között publikáltak, évente 12 számmal. A végleges korpuszunk összesen 13 185 200 tokent tartalmaz. A tokenek megoszlása a korpuszban kiegyensúlyozottnak tekinthető az egyes évek között, tehát a tokenek száma nagyjából megegyezik minden évben. Megjegyezzük azonban, hogy az 1956-os novemberi és decemberi

[3] https://github.com/tesseract-ocr

[4] http://www.inf.u-szeged.hu/rgai/magyarlanc

számok nem jelentek meg, ezért hiányoznak az összeállításunkból. A folyóirat utolsó száma 1989 áprilisában jelent meg.

4. A korpusz feldolgozása

Célunk a kiválasztott fogalmak szemantikai változásának feltárása volt, amihez a szóbeágyazás módszerét alkalmaztuk.

A szóbeágyazás alapvetően egy adott szótár szavainak vektorszerű ábrázolását jelenti, ahol a szóvektor dimenziójának alacsonyabbnak kell lennie, mint maga a szótár elemeinek a száma. A szótár egy adott dokumentumot vagy egy adott korpuszt reprezentál.

Az egyes nyelvi elemek vektorai alapján kiszámíthatjuk az egyes vektorok közötti távolságot, képet kapva ezáltal az adott két szó közötti szemantikai hasonlóságról, illetve különbségről. Egy adott beágyazási modellben ugyanis a hasonló kontextusban szereplő szavak vektorai hasonlóan helyezkednek ez az adott vektortérben, és a disztribúciós hipotézis alapján a szemantikailag hasonló szavak hasonló disztribúciós sajátságokkal (Harris, 1954), ezáltal pedig hasonló vektorreprezentációval rendelkeznek. Az elmondottakkal összefüggésben, a szóvektorok az időbeli szemantikai változások feltérképezésére is jól használhatóak. Amennyiben ugyanis a szavak vektorait különböző időszakokat reprezentáló korpuszok alapján készítjük el, azok összehasonlításával megkaphatjuk azok dinamikus változásait (Bamler és Mandt, 2017). A módszerrel többek között reprezentálhatóvá válhatnak egyes kulcsfogalmakat, illetve társadalmi csoportokat érintő változások az adott történelmi korszakok folyamatában (Hamilton és mtsai, 2016a; Garg és mtsai, 2018).

Mielőtt kiválasztottuk a jelen feladathoz legmegfelelőbb algoritmust, több megoldást is teszteltünk (Word2vec (Mikolov és mtsai, 2013a,b), FastText[5], GloVe (Pennington és mtsai, 2014)), amelyek közül a GloVe-t találtuk az adott vizsgálati cél szempontjból a legjobban működőnek. A módszerek kiértékelésekor elsősorban kvalitatív eszközökre támaszkodtunk és általunk kiválasztott kulcsfogalmak közelségét/távolságát vizsgáltuk. A Word2vec és a Glove hasonló eredményeket adott, a FastText betű alapú modellje, azonban történelmileg nagyon távol álló, de hasonló betűkből álló szavakat is egymáshoz közeli térbe helyezett el. A beágyazási módszerek és a tesztelési eljárások részleteinek tárgyalása e cikk keretein kívül esik, az alábbiakban a feldolgozási lépésekre összpontosítunk.

Annak céljából, hogy a feldolgozás számítási igényeit csökkentsük, első lépésben redukáltuk a vizsgálatba vont szavak számát: azokat, amelyek kevesebb mint ötször jelentek meg az a kiinduló szövegkorpuszban, töröltük.

Ezt követően a GloVe modellt a korpusz szavainak globális együtt-előfordulási statisztikája alapján tanítottuk. 10-es méretű ablakot használtunk az együttelőfordulási mátrix felépítéséhez, ami azt jelenti, hogy a célszó előtti és utáni 10 szót kezeltük a szó kontextusaként. 300 dimenziós beágyazási méretet választottunk, amely a legtöbb beágyazási algoritmus, köztük a word2vec és a GloVe

[5] https://fasttext.cc

alkalmazása esetében az alapértelmezett választás. Ennek megfelelően minden szót egy 300 valós számból álló vektor reprezentál. A tanításhoz az R-nyelvű text2vec csomagba (Selivanov és Wang, 2016) implementált GloVe algoritmust használtuk, ahol az iterációk maximális száma 10 volt.

A szavak hasonlóságát a szóvektorok koszinusz hasonlóságával számoltuk ki, ami a leggyakrabban használt metrika a beágyazáson alapuló elemzésekben. A maximális koszinusz hasonlóság 1, ami abban az esetben teljesül, ha két szóvektor orientációja teljesen azonos egymással, azaz pontosan ugyanabba az irányba mutatnak; 0, ha a vektorok merőlegesek egymásra; végül -1, ha a két vektor ellentétes irányba mutat, egymással 180 fokos szöget zár be.

A teljes korpuszt hat különböző, ugyanakkor részben átfedő időszakra osztottuk, ami fontos lépés volt az időbeli megközelítésünk szempontjából (Kozlowski és mtsai, 2018). Nyilvánvaló, hogy a szóbeágyazás minősége nagyban függ a korpusz minőségétől és méretétől. Mivel az időbeli változás tanulmányozása érdekében hat kisebb időszakra kellett felosztanunk az eredetileg viszonylag nagy méretű korpuszunkat, az egyes alkorpuszok mérete, amelyeken végül dolgoztunk, jelentősen kisebb volt. Ennek okán döntöttünk úgy, hogy az alkorpuszokban átfedő időszakokat is megengedünk. Fontos ugyanakkor hangsúlyozni, hogy, bár az időszakok átfedésben vannak, mindegyiknek megvan a saját, egymást nem átfedő vektortere, azaz minden egyes alkorpuszra készítettünk egy egyedi GloVe modellt. Az időszakok a következőek voltak: 1956-1965 (2 510 565 token), 1962-1968 (2 065 400 token), 1965-1972 (2 377 305 token), 1968-1976 (2 672 386 token), 1972-1982 (3 257 968 token), 1976-1989 (3 848 622 token).

Az alkorpuszok meghatározását követően minden időtartamra meghatároztuk ugyanannak a szónak az egyedi vektorát.

A beágyazások reprodukálhatóságát is teszteltük, a következőképpen: a tanítási folyamatot többször megismételtük egy-egy kiválasztott alkorpuszra, és csupán minimális eltéréseket tapasztaltunk a tesztelés eredményei között. Ugyanakkor úgy döntöttünk, hogy egy robusztus, statisztikailag megbízhatóbb eredmény elérése érdekében a következő megoldást alkalmazzuk: 20 beágyazási modellt készítünk minden időszakra, majd a 20 különálló vektor mindegyike esetében kiszámítjuk a kiválasztott fogalmakat reprezentáló vektorok közötti koszinusztávolságot. Végül, a 20 egyedi hasonlóság átlagával kapjuk a vizsgált fogalmak közötti tényleges hasonlósági mutatót (Antoniak és Mimno, 2018). Tesztjeink azt mutatták, hogy az alkalmazott megoldás az esetünkben stabil és megbízható eredményhez vezetett.

5. Eredmények

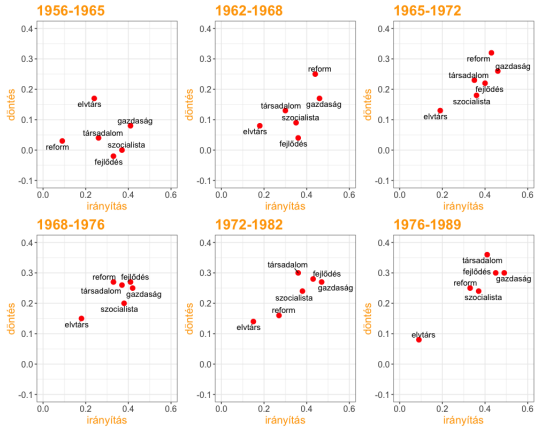
Elemzésünk során elvégeztük a vizsgált fogalmak gyakoriságának vizsgálatát, amelynek eredményét az alábbi szófelhők segítségével mutatjuk meg.

Vizsgálatunk jól mutatja a szavak gyakoriságának és ezen keresztül a korszak diskurzusában megjelenő kifejezések szerepének változását. Ezek közül kiemelendő a *döntés* szó gyakoriságának növekedése, amelynek következtében a húsz szó

2. ábra: A vizsgált szavak gyakorisága időszakonként

között az első időszak 19. helyéről a hatodik időszakban a 11. helyre lépett előre, míg az *elvtárs* szó az első periódus 6. helyéről a hatodik periódus 19. helyére esett vissza. Ha az első és az utolsó időszak leggyakoribb tíz-tíz kifejezését vetjük össze egymással, azt is kimutathatjuk, hogy az első tíz kifejezés egy kivételével ugyanaz, csak a szavak sorrendje változik. Egy lényeges kivételt tapasztalunk, az első időszakban még itt szereplő szavak közül kikerül az *elvtárs* kifejezés, ami az azonos politikai elveket vallókat jelöli és helyét a *társadalom* kifejezés veszi át, ami jól jelzi a diskurzus irányának megváltozását.

A következőkben a vizsgált fogalmak szemantikai összefüggéseit elemezzük. A 3. ábrán a fogalmak koszinusz közelségét mutatjuk meg az általunk meghatározott hat időszakban.



3. ábra: A vizsgált fogalmak koszinusz hasonlósága hat időszakban.

Eredményeink jól szemléltetik, hogy a *döntés* és *irányítás* szavak kapcsolata a vizsgált fogalmakkal jelentős változáson ment át. Az első két periódusban fogalmaink az *irányítás* szóhoz állnak közelebb, ami azt jelzi, hogy a diskurzus inkább direktívákat, mintsem alternatívákat is jelző döntési folyamatokra való utalásokat tartalmaz. Történetileg ez a magyar gazdaságpolitika azon időszaka, amikor a beruházások további finanszírozása és ezzel egyidejűleg az életszínvonal emelése komoly problémává vált. A források optimalizálása érdekében megindult a vállalatok összevonása, a „trösztösítés", az ország ipari vállalatait 15 nagyüzembe vonták össze. Felduzzadt a központi irányító apparátus, a helyi üzemegységek saját döntési lehetőségei viszont megszűntek.

Az 1960-as évek közepétől a kifejezések egyre közelebb kerülnek a *döntéshez*, ami a diskurzusban megjelenő alternatívákat jelzi. Az 1968-as reform, az „új gazdasági mechanizmus" értelmében nőtt az egyes vállalatok önállósága, a vállalatok saját megtermelt nyereségük egy részének beruházásáról maguk dönthettek. A mezőgazdaságban is növekedett a termelőszövetkezetek mozgástere, többek között engedélyezték számukra a jövedelmező melléküzemágak létesítését. A *társadalom* kifejezés például az első két időszakban meglehetősen gyenge kapcsolatban áll a *döntés* kifejezéssel (0.04-0.13), viszont erős a kapcsolata az *irányítással* (0.26- 0.30) a következő négy periódusban viszont, bár erős kapcsolata marad az *irányítással* (0.35-0.41), ugyanolyan erős kapcsolatba kerül a *döntéssel* (0.23-0.36)is. Érdekes tendenciát figyelhetünk meg a *reform* kifejezésnél is, amelynek a koszinusz közelsége az első időszakban mind a *döntéshez* (0.03), mind az *irányításhoz* (0.09) alacsony volt, a többi periódusban viszont magas, csak az 1970-es években tapasztalható a *reform-döntés* kapcsolatban némi gyengülés (0.16), ami magyarázható az 1968-as gazdasági reform ebben az időben történő leállításával. A *reform-irányítás* esetében minden időmetszetben 0.2 feletti küszöbértéket tapasztalhatunk. Szintén a korabeli gazdasági intézkedésekkel magyarázhatjuk a *gazdaság* kifejezéssel kapcsolatos eredményeinket. A *gazdaság* és az *irányítás* viszonyában minden vizsgált időszakban magas koszinusz küszöbértéket tapasztaltunk(0.41-0.49), míg a *döntéssel* csak az 1960-as évek második felétől. Ugyancsak figyelemre méltóak az *elvtárs* szóval kapcsolatos eredményeink, amely esetében minden periódusban mind a *döntéshez*, mind az *irányításhoz* viszonyítva 0.2 alatti küszöbértéket látunk.

6. Konklúzió

Dolgozatunkban a Kádár-korszak állampártjának hivatalos lapja alapján, történeti és szociológiai szempontok alapján kiválasztott kifejezések alapján az 1956 és 1989 közötti politikai diskurzus dinamikájának változásait vizsgáltuk. Elemzésünk fő célja az volt, hogy a korszak diskurzusának hosszanti, számítógépes és automatizált szövegelemzésére tegyünk kísérletet. Ehhez első lépésként összeállítottunk egy nagyméretű, 13 millió tokent tartalmazó, digitalizált korpuszt a *Pártélet* című folyóirat 379 számából. A korpusz nyers szövegeinek előfeldolgozását követően az adatokat szóbeágyazási módszerrel dolgoztuk fel. Ez a módszer lehetővé tette a vizsgált diskurzus néhány kulcsfogalma dinamikus változásainak

elemzését. Célunk az volt, hogy megvizsgáljuk a szóbeágyazás módszerének hasznosságát a történeti diskurzuselemzés számára. Elemzésünk megmutatta,hogy a hasonló korpuszok építése és kvantitatív elemzése hozzájárulhat a különféle történelmi korszakok diskurzusának mélyebb megértéséhez.

A jövőben célunk kutatásunkat más fogalmakra is kiterjeszteni, valamint ezen fogalmak dinamikus változásait is elemezni. A korpuszt további tisztítás és címkézés után terveink szerint elérhetővé tesszük a szélesebb kutatói közönségnek is. Végül, de nem utolsósorban, szeretnénk megvizsgálni a *Pártélet* és más célközönségnek szóló sajtótermékek diskurzusának dinamikai jellemzői közötti különbségeket és hasonlóságokat.

Köszönetnyilvánítás

A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM).

Hivatkozások

Antoniak, M., Mimno, D.: Evaluating the stability of embedding-based word similarities. Transactions of the Association for Computational Linguistics 6, 107–119 (2018)

Bamler, R., Mandt, S.: Dynamic word embeddings. In: ICML (2017)

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. In: Proceedings of the National Academy of Sciences of the United States of America (2018)

Gyáni, G.: A történelem mint emlék(mü). Kalligram, Budapest (2016)

Hamilton, W.L., Leskovec, J., Jurafsky, D.: Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing 2016, 2116–2121 (2016a)

Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. ArXiv abs/1605.09096 (2016b)

Harris, Z.S.: Distributional structure. WORD 10(2-3), 146–162 (1954), https://doi.org/10.1080/00437956.1954.11659520

Jatowt, A., Duh, K.: A framework for analyzing semantic change of words across time. In: IEEE/ACM Joint Conference on Digital Libraries. pp. 229–238 (2014)

Kozlowski, A.C., Taddy, M., Evans, J.A.: The geometry of culture: Analyzing meaning through word embeddings. ArXiv abs/1803.09288 (2018)

Kulkarni, V., Al-Rfou', R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. In: Proceedings of WWW (2014)

Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013a)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS (2013b)

Pap, M.: A népitöl a szocialista demokráciáig A korai Kádár-korszak demokráciafogalma a pártfolyóiratok tükrében. Múltunk 1, 202–226 (2017)

Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of EMNLP (2014)

Selivanov, D., Wang, Q.: text2vec: Modern text mining framework for r. Tech. rep. (2016), computer software manual](R package version 0.4. 0). Retrieved from https://CRAN. R-project. org/package= text2vec

Szabó, M.: A dolgozó mint állampolgár. Fogalomtörténeti tanulmány a magyar szocializmus három korszakáról. Korall 27, 151–171 (2007)

Xu, Y., Kemp, C.: A computational evaluation of two laws of semantic change. In: Proceedings of CogSci (2015)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP (2013)

Automatikus összefoglaló generálás magyar nyelvre BERT modellel

Yang Zijian Győző[1,2,3], Perlaki Attila[1], Laki László János[2,3]

[1]Eszterházy Károly Egyetem, Informatikai Kar
3300 Eger, Leányka út 4.
{yang.zijian.gyozo, perlaki.attila,}@uni-eszterhazy.hu
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[3]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/a.
{yang.zijian.gyozo, laki.laszlo}@itk.ppke.hu

Kivonat Cikkünkben különböző automatikus magyar nyelvű összefoglalást generáló neurális modelleket mutatunk be. Kétféle összefoglaló módszert különböztetünk meg. Az első módszer az absztraktív, amely a meglévő szövegből kinyeri a hasznos információt, majd erre támaszkodva próbál értelmes összefoglaló szöveget generálni. A másik módszer az extraktív, melynek lényege, hogy a meglévő szövegből azokat a mondatokat vagy kifejezéseket nyeri ki, amelyek leginkább leírják a szöveg tartalmi lényegét. A rendszer a kinyert szövegrészeket használja fel összefoglalóként. Kutatásunkban a „state-of-the-art" nyelvi reprezentációs modellnek számító BERT modellt használtuk. A rendszer tanításához különböző neurális modelleket alkalmaztunk. Extraktív összefoglaláshoz kipróbáltunk egy lineáris osztályozó, egy RNN és egy Transformer modellt. Az absztraktív modell tanításához Transformer modellt használtunk.
Kulcsszavak: extraktív összefoglaló, absztraktív összefoglaló, BERT

1. Bevezetés

A nagy mennyiségű írott szövegek rendszerezéséhez és átláthatóságához elengedhetetlen azok kivonatolása. Erre napjainkban automatikus rendszerek léteznek, melyek feladata a hosszabb szövegek, szövegrészek összefoglalása – text summarization – oly módon, hogy önálló folyékony szövegként leírja az egész dokumentum lényegét.

Kétféle automatikus összefoglaló megközelítést különböztetünk meg: absztraktív és extraktív. Absztraktív szöveg-összefoglalásnak hívjuk azt, amikor egy szöveg alapján olyan szöveget generálunk, amely kivonata az eredeti szövegnek. Tartalmazza a lényeget, tömörebben, rövidebben fogalmazza meg azt. Ez a módszer hasonlít leginkább az emberi összefoglaláshoz. A módszer legnagyobb nehézsége, hogy olyan szöveget kell generálni, ami nemcsak nyelvileg helyes, hanem tartalmaznia kell az eredeti szöveg mondanivalóját is. Ez két igen nehéz feladat, amelyekből külön-külön is több kutatás folyik. A másik megközelítés az

extraktív összefoglaló generálás, amely az eredeti szövegből nyeri ki a lényegre vonatkozó szövegrészleteket. Ez a feladat annyiban könnyebb, hogy nem kell nyelvileg helyes mondatot generálni, elég csak a meglévő szövegből megkeresni a lényegre vonatkozó részeket.

Kutatásunk célja, hogy megvizsgálja a jelenlegi legjobb eredményt elért összefoglaló módszert magyar nyelvre. Továbbá szeretnénk egy működő magyar nyelvű absztraktív és extraktív összefoglaló rendszert létrehozni.

2. Kapcsolódó irodalom

Az extraktív módszer a legfontosabbnak ítélt mondatok kiemelésével (és szükség szerinti egyesítésével) dolgozik, a neurális modell szempontjából ez osztályozási problémaként jelenik meg: mely mondatok választhatók ki arra, hogy az összefoglalóban is szerepeljenek. Az egyik legkorábbi neurális hálózaton alapuló extraktív rendszer a SummaRuNNer (Nallapati és mtsai, 2017), amely egy RNN enkóder segítségével oldja meg a problémát. A Refresh (Narayan és mtsai, 2018) Rouge metrikán alapul, melynek segítségével megerősítéses tanulásos módszerrel rangsorolják a mondatokat a szövegben. A Latent (Zhang és mtsai, 2018) célja a kulcsszavak legpontosabb követése helyett az emberi munkával készült absztraktokhoz való minél közelebbi hasonlóság elérése volt. A Sumo (Liu és mtsai, 2019) olyan módszert alkalmaz, amely a dokumentumból kinyerhető több-gyökerű függőségi fa-struktúrákra épül, és az összefoglaló lehetséges formájának előbecslésén alapszik. A NeuSum (Zhou és mtsai, 2018) a mondatok pontozásával és szelektálásával közelíti meg a problémát.

Az absztraktív módszer neurális megközelítésben olyan problémaként mutatkozik meg, ahol egy adott szekvenciát egy másik szekvenciába kell transzformálni. Az enkóder a forrás-dokumentumból tokeneket azonosít be, azokat feltérképezi, majd a dekóder tokenről tokenre állít elő ebből egy új szöveget. A PTgen (See és mtsai, 2017) egy mutatókat (pointer) generáló eszköz, amely a forrásszövegben szavakat azonosít be, ezután egy közvetítő (coverage) mechanizmus az összefoglalóba kerülő szavakat tartja meg. A Deep Communicating Agent (Celikyilmaz és mtsai, 2018) olyan ágens-alapú megközelítés, ahol az ágensek együtt reprezentálják a feldolgozandó dokumentumot és ennek dekódolásákoz kapcsolódik egy hierarchia-figyelő ágens. A Deep Reinforced Modell (Paulus és mtsai, 2018) közvetítés-alapú (coverage), ahol a dekóder a már generálásra került szöveget is figyeli. A BottomUp (Gehrmann és mtsai, 2018) tartalomszűrő eljárása előbb meghatározza, mely szövegrészek kerülhetnek az összefoglalóba, majd a dekóder már csak ezeken dolgozik.

Magyar nyelven az OpinHu rendszer (Miháltz, 2010) rendelkezik összefoglaló funkcióval. A rendszer kulcsszavakat és szövegkontextust használ az információkinyerésre.

3. Az összefoglaló rendszer

Ebben a fejezetben mutatjuk be az összefoglaló rendszer részeit és a mögötte lévő korpuszt. Továbbá megismertetjük a BERT modell architektúráját, valamint az ezen alapuló absztraktív és extraktív modelleket.

3.1. A BERT modell

Yang Liu és Mirella Lapata szöveg-összefoglalással kapcsolatos munkája (Liu és Lapata, 2019) az előtanított nyelvi modellek (ELMo, GPT, BERT) közül a BERT modellt (lásd 1. ábra bal oldala) emeli ki. Ez a modell rendelkezik szó-, mondat- és pozícióreprezentációval is, amely nagyméretű szövegkorpuszon alapszik. A legtöbb esetben az előtanított modellek olyan természetes nyelvi feldolgozási problémák esetén alkalmazhatók, ahol mondat- illetve bekezdés-szintű értelmezés, osztályozás szükséges. Cikkükben bemutatják, hogy a szöveg-összefoglalás feladata túlmutat az egyszerű szó- vagy mondatfordításon.

A BERT (Devlin és mtsai, 2019) modell egy előre tanított nyelvi reprezentáció, a „Bidirectional Encoder Representations from Transformers" rövidítése, a Google terméke. A BERT modell tanítása két lépésből áll: „pre-training" és „fine-tuning". A „pre-training" fázisban egy általános nyelvi reprezentációt tanítanak, majd ezen modell kimeneti paramétereinek segítségével a „fine-tuning" fázisban egy feladatspecifikus modellt tanítanak be. A BERT modell tanítása úgy történik, hogy a szövegből először WordPiece (Wu és mtsai, 2016) tokenizálóval egy általános nyelvfüggetlen szótárat hoznak létre, majd a tokenizált szöveg véletlenszerűen kiválasztott 15%-át elmaszkolják, végül a modell ezeket az elmaszkolt szövegrészeket próbálja kitalálni. Ezután végeznek egy becslést a következő mondatra, melyből 50% valódi és 50% véletlenszerű mondat. A tanításhoz kétirányú Transformer modellt (Vaswani és mtsai, 2017) használnak.

A Google betanított két többnyelvű modellt is[1]: kisbetűsített és nem kisbetűsített. A modellek tanításához kiválasztották az első 104 nyelvet, amely a legnagyobb Wikipédiával rendelkezik. A egyes nyelvek Wikipédia mérete igen különbözik, az adat közel 20%-át teszi ki az angol Wikipédia, ezért normalizálással kontrollálták a mintavételezést, hogy kiküszöböljék ezt a problémát. Ezután minden nyelvet, hasonlóan az angolhoz, tokenizálásnak vetették alá, amelynek négy lépése volt: kisbetűsítés, ékezetek eltávolítása, írásjelek leválasztása, whitespacek kezelése. A nem kisbetűsített modell tanítása is ezeken a lépéseken esett át, a WordPiece szótár segítségével kezelik a nem kisbetűs és ékezetes szavakat.

Természetesen a magyar nyelv is része ennek a modellnek. Kutatásunkhoz a nem kisbetűsített többnyelvű modellt (BERT-Base, Multilingual Cased) használtuk.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

3.2. A korpusz

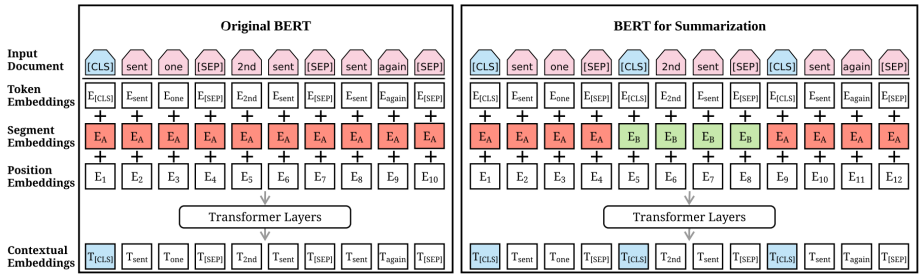
Tanító- és tesztkorpuszként a hvg.hu által rendelkezésünkre bocsátott nyomtatott és online hírlapból vett cikkeket, valamint a hozzájuk tartozó leadeket használtuk fel. A korpusz tulajdonságai:

- Nyomtatott cikkadatbázis (hetilap): 1994-2017
  - 35.513 cikk; 34.409.106 token; 2.045.255 type
- Online cikkadatbázis (napilap): 2012-2017
  - 374.064 cikk; 87.366.132 token; 3.544.622 type
- Összesen: 346.873 cikk; 121.772.523 token; 4.365.813 type;
- Cikkek témái: gazdaság, politika, tudomány, sport, kultúra, pszichológia, blog
- Kísérlethez:
  - Tanítóanyag: 343.000 cikk
  - Tesztanyag: 1790 cikk (eredtileg 1873 cikk volt, csak a rendszer kivette azokat a cikkeket, amelyek háromnál kevesebb mondattal rendelkeztek)
  - Validálás: 2000 cikk
  - Forrásszöveg (cikkek) átlagos bekezdéshossza: 317,37 szó; 15,36 mondat
  - Célszöveg (lead) átlagos bekezdéshossza: 26,21 szó; 1,56 mondat

Mivel nem mindegyik cikkhez tartozott lead, ezért miután a nyomtatott és az online cikkeket összetettük, kivettük azokat a cikkeket, amelyekhez nem tartozott lead (ezért látható eltérés a tulajdonságban az Összesen résznél).

3.3. Az extraktív és az absztraktív modellek

A BERT modell hangolható („fine-tune”) más feladatokra is, mint például összefoglaló generálásra. Az összefoglaló generáláshoz az összefoglalókból (a mi esetünkben leadekből) képzett szegmensvektorok (mondatvektorok) bemenetként szolgálnak az egyes Transformer rétegek számára. Kétféle összefoglaló modellt tudunk behangolni így: extraktív és absztraktív modellek.
Az extraktív modell: A BERT modell kimenetére rákötnek egy plusz réteget, ami specifikus feladatra alkalmas. A mi esetünkben egy olyan réteget kötnek rá, mely segítségével a cikkben lévő minden egyes mondatra mond egy valószínűséget, hogy az milyen mértékben írja le a cikk tartalmi lényegét. Az, hogy egy mondat mennyire foglalja össze a cikket, a leadeket használja fel. A plusz réteg lehet egyrészt egy sima lineáris osztályozó szigmoid függvénnyel, másrészt egy LSTM architektúrájú RNN, harmadrészt egy Transformer modell. A rendszer a betanított modellel kiválasztja és rangsorolja a cikkből azt a 3 mondatot, ami leginkább leírja annak tartalmi lényeget.
Az absztraktív modell: Az absztraktív modell megfeleltethető egy enkóder-dekóder alapú gépifordító rendszernek, ahol a forrásnyelv maga a dokumentum, a célnyelv pedig annak összefoglalója. A tanításhoz ebben az esetben a forrásnyelvi oldalon a BERT modellt használjuk, míg a célnyelvi oldalon a tanító anyagunk leadjeit használjuk.



1. ábra: A BERT és az összefoglaló BERT architektúrája (Liu és Lapata, 2019)

4. Kísérletek

Első lépésünk az volt, hogy előfeldolgozást végeztünk az eredeti szövegeken, mely az alábbi lépésekből állt. A cikkeket először mondatokra bontottuk, majd tokenizáltuk. Ezekhez az e-magyar tokenizálóját, a quntoken (Mittelholcz, 2017) eszközt használtuk. Ezt követően a szöveget az összefoglaló rendszer számára JSON formátumra alakítottuk. A rendszer ezután két speciális elemet illeszt be, az egyik a szöveg elejét jelzi, a másik a mondathatárokat. Ezután az előfeldolgozott fájlokkal különböző neurális modelleket tanítottunk be.

Kutatásunkban először megmértük az alapmódszer (baseline) teljesítményét, amely a cikk első három mondatát veszi összefoglalóként.

Következő lépésként betanítottunk három modellt az extraktív összefoglalóhoz:

- Lineáris osztályozó (BERT-Class), ahol a BERT modell kimenetére egy szigmoid függvénnyel ellátott lineáris réteg van kötve.
- Rekurrens neurális modell (BERT-RNN), melyben a BERT modell kimenetére egy bidirekciós LSTM réteg van kötve.
- Transformem modell (BERT-TransExt), ahol a BERT modell kimenetére egy Transformer modell van kötve.

Az absztraktív összefoglalóhoz kipróbáltunk két modellt:

- Baseline Transformer modell (BERT-TransAbs): egy alapértelmezett Transformer modell.
- Baseline absztraktív Transformer modell (BERT-TransAbs-baseline): Yang Liu és Mirella Lapata kutatásában (Liu és Lapata, 2019) az absztraktív modellre behangolt („fine-tuned”) baseline modell.

A modellek tanításhoz a Yang és társa (Liu és Lapata, 2019; Liu, 2019) által implementált eszközöket[2] használtuk fel.

A beállítási paraméterek extraktív modellek esetén:

[2] https://github.com/nlpyang

– Általános paraméterek:
  • dropout: 0,1; learning rate: 2e-3; batch size: 3000; tanítási lépésszám: 100000
– Transformer modell:
  • head: 8; belső réteg: 2; feedforward méret: 2048
– RNN modell:
  • rnn méret: 768

A beállítási paraméterek absztraktív modell esetén:

– dropout: 0,1; learning rate: 0,05; batch size: 3000; tanítási lépésszám: 200000; rejtett rétegek neuronszáma (enkóder, dekóder): 512; rétegek száma (enkóder, dekóder): 6; feedforward méret (enkóder, dekóder): 2048

5. Eredmények

A kiértékeléshez a ROUGE (Lin, 2004) módszert használtuk. A ROUGE (Recall-Oriented Understudy for Gisting Evaluation) egy fedés alapú módszer, ami a gépi fordítás során használt BLEU metrikán alapszik. Maga a ROUGE több almetódust is tartalmaz, melyek közül a méréseinkhez a ROUGE-1, ROUGE-2 és a ROUGE-L módszereket használtuk. A ROUGE-1 egy unigram, míg a ROUGE-2 egy bigram fedést számoló algoritmus. A ROUGE-L a leghosszabb közös szósorozatot vizsgálja bekezdés és mondat szinten.

A 1. táblázatban láthatók a különböző modellek teljesítményei. Az alapmódszer (baseline) eredménye teljesített a leggyengébben. Extraktív modell esetén BERT-RNN modell érte el a legjobb eredményt. Itt érdemes megjegyezni, hogy angol nyelv esetében ez a Transformer modell volt. Az eredmények csak azt mutatják, hogy az gép által kiválasztott mondat mennyire hasonlít a leadre. Lehetséges problémaforrás, hogy sok esetben a leadnek nem összefoglaló, hanem figyelemfelkeltő szerepe van.

Az absztraktív modell eredményeit tekintve igen alacsony a fedés, ami önmagában csak annyit jelent, hogy az összefoglaló nem hasonlít a leadre, de a kimenetet nézve sajnos egyelőre nem tudjuk értékelni még ezeket az eredményeket, mert a rendszer túltanult és mindenre ugyanazt a mondatot adta eredményül. A továbbiakban csak az extraktív modelleket fogjuk elemezni.

A 2. táblázatban látható az extraktív modellek viselkedése egymáshoz viszonyítva. Látható, hogy az esetek közel 7%-ában pontosan ugyanabban a sorrendben ajánlották ugyanazokat a mondatokat összefoglalásnak. Továbbá az látható, hogy a 3 modell közel 30%-ban ugyanazt a 2 mondatot választotta ki, és szintén közel 30%-ban pontosan egy közös mondatot választottak. Az arányokat nézve nagyon ritka eset az, amikor nem volt közös mondat. A páros összehasonlításokat nézve az szembetűnő, hogy az RNN és az osztályozó modell sokkal hasonlóbban választottak mondatokat, mint a Transformer és az osztályozó modell.

Az egyik legalapvetőbb extraktív összefoglaló módszer az, hogy kiválasztjuk a forrásszöveg első néhány mondatát (Liu és Lapata, 2019). A 3. táblázatban láthatjuk azokat az eredményeket, amelyek azt mutatják, hogy a különböző modellek milyen arányba választották a forrásszöveg első három mondatát. A rendszer

Model	ROUGE-1	ROUGE-2	ROUGE-L
Extraktív			
baseline	54,58	27,25	45,52
BERT-Class	55,26	28,21	46,23
BERT-RNN	**55,46**	**28,29**	**46,27**
BERT-TransExt	54,76	27,71	45,97
Absztraktív			
BERT-TransAbs	27,73	2,89	23,85
BERT-TransAbs-baseline	16,04	1,36	13,72

1. táblázat. ROUGE fedés eredmények

	Egyező mondatok száma		
	3 db	2 db	1 db
BERT-RNN - BERT-Class	33,18%	46,42%	12,35%
BERT-RNN - BERT-Trans	20,95%	42,35%	23,85%
BERT-Trans - BERT-Class	20,89%	32,35%	24,08%
BERT-RNN - BERT-Trans - BERT-Class	13,02%	35,92%	32,12%
BERT-RNN - BERT-Trans - BERT-Class (sorrend is egyezik)	6,93%		

2. táblázat. A különböző extraktív modellek viselkedése egymáshoz viszonyítva

a forrásszövegből rangsorolva 3 mondatot ajánl összefoglalónak. Az eredményből azt láthatjuk, hogy a Transformer modell első ajánlásnak közel 80%-ában választ a forrásszöveg első három mondatából, az esetek felében az első mondatot választja ki annak. Másik kiemelkedő eredmény az RNN modell viselkedése, amely közel 72%-ban választja a forrásszöveg első mondatát valamelyik ajánlásnak. Az esetek közel 40%-ában választja az első mondatot első ajánlásnak.

A 4. táblázatban látható néhány példa a különböző modellek kimeneteire. Láthatunk először példát arra, amikor teljesen megegyezik mind az ajánlott mondatok, mind a sorrend (a 2. táblázat alapján az esetek 6,93%-a). Majd mutatunk példát arra, amikor az ajánlott mondatok megegyeznek, de más sorrendben ajánlanak (a 2. táblázat alapján az esetek 13,02%-a). Ezután láthatunk néhány példát arra, amikor közel hasonló eredményeket adtak a különböző modellek. A példában a BERT-Class és a BERT-RNN modellek ugyanazokat a mondatokat ajánlották, csak más sorrendben (a 2. táblázat alapján az esetek 33,18%-a). A Transformer modell harmadik ajánlása különbözik a másik kettő modellétől. Végül egy olyan példát lehet látni, ahol eléggé különböző ajánlásokat adtak a modellek, a példában egy közös mondat van csak.

6. Összegzés

Létrehoztunk egy magyar nyelvű szöveg-összefoglaló rendszert, amellyel jelenleg extraktív összefoglalást tudunk készíteni hírlap cikkekből. A rendszer tanításhoz a jelenleg „state-of-the-art” nyelvi reprezentáció modellt, a Google által kuta-

	1. ajánlás	2. ajánlás	3. ajánlás	Összesen
	1. mondata a forrásszövegnek			
BERT-Class	38,60%	18,83%	11,56%	68,99%
BERT-RNN	40,89%	18,27%	12,79%	71,96%
BERT-Trans	52,46%	7,04%	6,48%	65,98%
	2. mondata a forrásszövegnek			
BERT-Class	17,21%	28,38%	15,53%	61,12%
BERT-RNN	19,05%	27,21%	15,92%	62,18%
BERT-Trans	16,65%	41,62%	6,76%	65,03%
	3. mondata a forrásszövegnek			
BERT-Class	11,90%	14,08%	24,36%	50,34%
BERT-RNN	11,73%	15,59%	22,07%	49,39%
BERT-Trans	11,28%	17,65%	42,23%	71,17%
	Összesen			
BERT-Class	67,71%	61,28%	51,45%	
BERT-RNN	71,68%	61,06%	50,78%	
BERT-Trans	80,39%	66,31%	55,47%	

3. táblázat. A forrásszöveg első három mondatának kiválasztásának arányai

tott többnyelvű BERT modellt használtuk. Az extraktív összefoglalóhoz többféle modellt is kipróbáltunk, melyek közül az RNN érte el az legjobb eredményt. Az absztraktív összefoglaláshoz Transformer alapú neurális hálót használtunk. Sajnos az absztraktív modellünk még nem ért el értékelhető eredményt, de az extraktív modellek már működnek és eredményeinkben kielemeztük működéseit. Továbblépésként az absztraktív modellekkel szeretnénk értékelhető eredményt elérni.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

A kutatást az EFOP-3.6.1-16-2016-00001 „Kutatási kapacitások és szolgáltatások komplex fejlesztése az Eszterházy Károly Egyetemen” című projekt támogatta.

Hivatkozások

Celikyilmaz, A., Bosselut, A., He, X., Choi, Y.: Deep communicating agents for abstractive summarization. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1662–1675. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)

Modell	Példa
BERT-Class BERT-RNN BERT-Trans	1. Pótlóbuszok járnak az M3-as metró helyett Újpest-Központ... 2. Mintegy háromnegyed órával később közölték: helyreállt a rend... 3. Kérjük az arra közlekedők türelmét – írták a BKK....
BERT-Class	1. A napokban Tajvanról érkezett egy sisakkamerás felvétel... 2. Az incidenst egy motoros társaság közös csapatása során... 3. Jó kérdés, hogy mit szólt a barátnő a férfi magatartásához...
BERT-RNN	1. Az incidenst egy motoros társaság közös csapatása során... 2. Jó kérdés, hogy mit szólt a barátnő a férfi magatartásához... 3. A napokban Tajvanról érkezett egy sisakkamerás felvétel...
BERT-Trans	1. Jó kérdés, hogy mit szólt a barátnő a férfi magatartásához... 2. Az incidenst egy motoros társaság közös csapatása során... 3. A napokban Tajvanról érkezett egy sisakkamerás felvétel...
BERT-Class	1. A finn kormányfő ugyanakkor meg van győződve arról... 2. A finn kormány 2017 januárjától próbaképpen bevezetné... 3. A kormányfő az intézkedéstől a szociális juttatások rendszerének...
BERT-RNN	1. A finn kormány 2017 januárjától próbaképpen bevezetné... 2. A kormányfő az intézkedéstől a szociális juttatások rendszerének... 3. A finn kormányfő ugyanakkor meg van győződve arról...
BERT-Trans	1. A finn kormány 2017 januárjától próbaképpen bevezetné... 2. A kormányfő az intézkedéstől a szociális juttatások rendszerének... 3. A társadalmi kísérlet a 2015-ben hivatalba lépett...
BERT-Class	1. Az óceánparti San Sebastián baszk nagyvárosban az óvárosig... 2. Megrongálódott három híd, amely az Urumea folyón vezetett át. 3. A létesítmény igazgatója több mint kétmillióra euróra tette a kárt.
BERT-RNN	1. Zarauzban és az északspanyol part más fürdőhelyein épületek... 2. Megrongálódott három híd , amely az Urumea folyón vezetett át. 3. Az óceánparti San Sebastián baszk nagyvárosban az óvárosig...
BERT-Trans	1. Asztúria autonóm körzetben a hullámok lerombolták a luarcai... 2. Az óceánparti San Sebastián baszk nagyvárosban az óvárosig... 3. Mint az elpais.com, az El País című lap internetes portálja...

4. táblázat. Néhány példa az extraktív modellek kimeneteire

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

Gehrmann, S., Deng, Y., Rush, A.: Bottom-up abstractive summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4098–4109. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)

Liu, Y.: Fine-tune bert for extractive summarization. In: IJCNLP. Hong Kong, China (2019)

Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: IJCNLP. Hong Kong, China (2019)

Liu, Y., Titov, I., Lapata, M.: Single document summarization as tree induction. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1745–1755. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

Miháltz, M.: Opinhu: online szövegek többnyelv véleményelemzése. VII. Magyar Számítógépes Nyelvészeti Konferencia pp. 14–23 (2010)

Mittelholcz, I.: emtoken: Unicode-képes tokenizáló magyar nyelvre. XIII. Magyar Számítógépes Nyelvészeti Konferencia pp. 61–69 (2017)

Nallapati, R., Zhai, F., Zhou, B.: Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 3075–3081. AAAI'17, AAAI Press (2017)

Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization with reinforcement learning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1747–1759. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)

Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada (2018)

See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1073–1083. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett,

R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. Technical Report abs/1609.08144 (2016)

Zhang, X., Lapata, M., Wei, F., Zhou, M.: Neural latent extractive document summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 779–784. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., Zhao, T.: Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 654–663. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Korpusznyelvészet, szintaxis

1956 és 1989 között keletkezett propagandaszövegek nyelvi sajátságai

Szabó Martina Katalin[1,2], Ring Orsolya[1], Vincze Veronika[3]

[1] Társadalomtudományi Kutatóközpont, CSS-RECENS Kutatócsoport
1097 Budapest, Tóth Kálmán u. 4.
[2]Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.
[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.
ring.orsolya@tk.mta.hu, {martina,vinczev}@inf.u-szeged.hu

Kivonat Elemzésünkben a Magyar Szocialista Munkáspárt Központi Bizottságának (MSZMP KB) hivatalos havi lapja, a *Pártélet* első és utolsó két évének lapszámait vizsgáljuk, és összehasonlítjuk a két időszak szövegeinek statisztikai, morfológiai, szintaktikai és szemantikai jellemzőit. A vizsgált szövegek bizonyosan a totalitárius nyelvhasználatot reprezentálják, ezért kutatási eredményeinket felhasználhatónak tekintjük a politikai propaganda azonosítására és elemzésére irányuló kutatásokban és fejlesztésekben.
Kulcsszavak: információkinyerés, propaganda, megtévesztés, diskurzuselemzés, szocializmus, politika

1. Bevezetés

A magyar történelem 1956 és 1989 közötti időszaka a történettudományban és a társadalomtudományban a gyakran vizsgált korszakok közé tartozik. Politikai diskurzusának nyelvi jellemzői azonban eddig nem képezték elemzés tárgyát. Ezt a problémát felismerve korpuszt építettünk a *Pártélet* nevű folyóiratból, az MSZMP KB hivatalos ideológiai lapjából, majd azt számítógépes nyelvészeti eszközökkel és módszerekkel elemeztük.

A *Pártélet* című lap vizsgálata lehetővé teszi az országot irányító állampárt hivatalos diskurzusának tüzetes elemzését. A lap célja a politikai ideológia terjesztése, tehát a közvetlen agitáció és propaganda volt. A *Pártélet*, amely 54 150 példányban jelent meg, elsősorban nem az átlagemberekhez, hanem az állampárt különböző tisztségviselőihez szólt. Bár nincsenek pontos adataink arra vonatkozóan, hogy pontosan kik olvasták, a nagy példányszámból arra következtethetünk, hogy a kiadvány valószínűsíthetően a párthierarchia egészét célozhatta. Dolgozatunkban két, megközelítőleg egyforma hosszúságú időszakot vizsgálunk: a 1957 januárja és 1958 decembere közötti (17 szám), valamint a 1988 januárja és 1989 áprilisa közötti időszakot (16 szám). A két időszakot a következő szempontok alapján jelöltük ki: Az első időszak a Kádár-korszak kezdete, közvetlenül az 1956-os forradalom után következő, hipotézisünk szerint az agitáció

és a propaganda szempontjából igen aktív periódus. A második időszakot, amely közvetlenül megelőzi a rendszerváltást és a Kádár-korszak végét jelenti, az első ellenpontjának tekinthetjük. Az első a rendszer megszilárdulásának, míg a második annak fokozatos felbomlásának időszaka, így azt feltételezzük, hogy a kettő diskurzusa eltérést mutat.

Célunk a propaganda nyelvi jellemzőinek feltárása volt, különös tekintettel a politikai nyelvhasználatot gyakorta jellemző előítéletekre és a megtévesztésre (Propaganda Analysis, 1938; Jalilifar és Alavi, 2011; Barrón-Cedeño és mtsai, 2019). Azt vizsgáljuk, hogy a politikai célok miként és milyen mértékben befolyásolják a nyelvhasználatot, a diskurzus szervezését érintő döntéseket (Jalilifar és Alavi, 2011). Elemzésünk egyben egy jövőbeli, szélesebb körű – a propagandát az online térben vizsgáló – kutatás megalapozásának tekinthető.

Annak ellenére, hogy a megtévesztés és elfogultság jelenségei különböző típusú diskurzusokban azonosíthatóak, a legtöbb szerző ezeket vagy hangzó szövegekben vizsgálja (Fetzer, 2008; Fraser, 2010; Scheithauer, 2007; Simon-Vandenbergen és mtsai, 2007) vagy automatikus ténykivonatolási és bináris osztályozási feladatként kezeli (Greene és Resnik, 2009; Rubin és mtsai, 2015; Wang, 2017; Thorne és Vlachos, 2018; Graves, 2018). Bár a propaganda mint diskurzus kvantitatív és kvalitatív elemzése mind a számítógépes nyelvészet, mind a történettudomány és a politológia szempontjából fontos terület, eddig nagyon kevés kísérlet történt nagy mennyiségű propagandaszöveg szisztematikus elemzésére (Propaganda Analysis, 1938; Rashkin és mtsai, 2017; Barrón-Cedeño és mtsai, 2019). Különösen jelentős a hiány a magyar nyelvű szövegek vonatkozásában. Olyan dolgozatról pedig egyáltalán nincs tudomásunk, amely a magyar totalitárius diskurzus jellemzőit NLP-módszerekkel vizsgálná.

2. A vizsgálati korpusz

A vizsgálatokhoz használt korpuszt, a Pártélet című lap számait az Arcanum Digitheca[1] oldalról töltöttük le. A lap szkennelt, PDF-formátumban közzétett dokumentumait a letöltés után további komplex feldolgozási folyamatoknak vetettük alá, amelyek eredményeképpen megkaptuk a szövegek elemezhető és megfelelő minőségű, szöveges (txt) formátumú változatát (Szabó és mtsai, 2019).

A kész korpuszból a jelen dolgozatban bemutatott elemzésekhez kiválasztottuk a korpusz első két, és utolsó két évnyi lapszámát. E két alkorpusz alapvető statisztikai adatait az 1. táblázat közli.

korpusz	**lapszám**	**mondatszám**	**tokenszám**
1956-57	17	23573	503047
1988-89	16	28229	531962
Összesen	33	51802	1035009

1. táblázat. Az alkorpuszok alapvető adatai.

[1] https://adtplus.arcanum.hu/en/collection/Partelet/

Az alkorpuszok kiegyenlített mennyiségi adatai elősegítik az alkorpuszokban mért adatok összevetését egymással.

3. A vizsgálat módszere

Kutatásunk kiindulópontja az a hipotézis, hogy a *Pártélet* szerzői a megtévesztés és manipuláció különböző nyelvi eszközeit alkalmazták arra, hogy elrejtsék az akkori politikai vezetés számára nem kívánatos tényeket, és egyben befolyásolják, meggyőzzék a szövegek olvasóit (Girlea és mtsai, 2016; Rashkin és mtsai, 2017). Mivel elemzésünk alapja egy ideológiai folyóirat, amely a párt egyik fő szócsöve volt, azt feltételeztük, hogy valószínűleg a fenti jellemzők mindkét időszakban meghatározóak voltak, és az első időszakban elsősorban a forradalomhoz, a másodikban pedig a rendszerváltáshoz kapcsolódtak. Ugyanakkor azt is feltételeztük, hogy a két alkorpusz nyelvi jellemzői több sajátság tekintetében is különböznek egymástól, lévén, hogy az első alkorpusz szövegei a rendszer megerősödésekor, míg a második alkorpusz szövegei a rendszer gyengülésekor keletkeztek.

Első lépésben a szövegekre lefuttattuk a magyarlanc nyelvi elemzőt, így megkaptuk a szövegek mondat- és tokenszámát, lemmatizált változatát, majd morfológiai és szintaktikai elemzését (Zsibrita és mtsai, 2013). Ezt követően a szövegek szemantikai és pragmatikai sajátságait különböző szótárak segítségével elemeztük, például a nyelvi bizonytalanságot jelölő elemek (Vincze, 2014), a szentimentek és az emóció szótáraival (Szabó és mtsai, 2016; Szabó, 2015), és kiszámoltuk az egyes lexikális elemek számát és gyakoriságát. Az automatikus elemzés alapján kiszámított nyelvi jellemzőket az alábbiakban mutatjuk be.

- **Statisztikai jellemzők**:
  - mondatok száma,
  - szavak száma,
  - lemmák száma,
  - mondatok átlagos hossza.
- **Morfológiai jellemzők**:
  - főnevek, igék, melléknevek, határozószavak, tulajdonnevek, névmások, számnevek, kötőszavak, írásjelek és ismeretlen szavak száma és aránya a szószámhoz képest,
  - múlt és jelen idejű igék száma és aránya az igék számához képest,
  - feltételes és felszólító módú igék száma és aránya az igék számához képest,
  - műveltető, ható és gyakorító igék száma és aránya az igék számához képest,
  - E/1. és T/1. igék száma és aránya az igék számához képest,
  - mutató névmások száma és aránya a szavak számához képest,
  - felsőfokú és középfokú melléknevek száma és aránya a melléknevek számához képest.
- **Szintaktikai jellemzők**:
  - alanyok, tárgyak, jelzők, határozók, alárendelő és mellérendelő mondatok száma és aránya.

– **Szemantikai jellemzők**:
  - tagadószavak száma és aránya a szavak számához képest,
  - tartalmas szavak és funkciószavak aránya,
  - pozitív és negatív jelentésű szavak száma és aránya a szószámhoz képest (a listákat Szabó (2015) alapján készítettük el),
  - bizonytalanságra utaló szavak száma és aránya a szószámhoz képest (Vincze (2014) alapján felállított osztályokba sorolva),
  - érzelmekre utaló szavak száma és aránya a szószámhoz képest (Szabó és mtsai (2016) alapján felállított osztályokba sorolva),
  - befelé forduló cselekvésre utaló igék („private verbs") száma és aránya Quirk és mtsai (1985) alapján magyarítva,
  - kifelé forduló cselekvésre utaló igék („public verbs") száma és aránya Quirk és mtsai (1985) alapján magyarítva,
  - érvelésre utaló igék („suasive verbs") száma és aránya Quirk és mtsai (1985) alapján magyarítva.

  A szemantikai jellemzők vizsgálatakor egyszerű listaillesztéses módszert használtunk: amennyiben az egyes szavak lemmája megegyezett bármelyik listaelemmel, akkor találatként számoltuk.
– **Pragmatikai jellemzők**:
  - beszédaktust jelentő igék száma és aránya,
  - diskurzusjelölők száma és aránya Dér és Markó (2007) alapján,
  - szó szerinti idézetek és nyilatkozatok száma és aránya.

  Utóbbiakat a szövegben előforduló idézőjelek és mondatkezdő gondolatjelek számával mértük.

A két alkorpuszon kapott elemzési eredményeket végül összevetettük egymással statisztikai szignifikanciavizsgálatok (t-próba) segítségével. Eredményeinket összehasonlítottuk a Propagandaelemző Intézet (Institute for Propaganda Analysis) kutatási eredményeivel is (Propaganda Analysis, 1938).

Annak céljából, hogy a diskurzusban megjelenő időbeli változást még behatóbban vizsgálhassuk, a gyakorisági jellemzőket lapszámonként számítottuk ki, és eszerint ábrázoljuk majd azokat az eredményeket tárgyaló fejezetben is.

4. Eredmények

Vizsgálataink több sajátság tekintetében szignifikáns eltérést mutattak a két alkorpusz között, l. a 2. táblázat. Az alábbiakban bemutatunk néhányat a leginkább figyelemre méltóak közül.

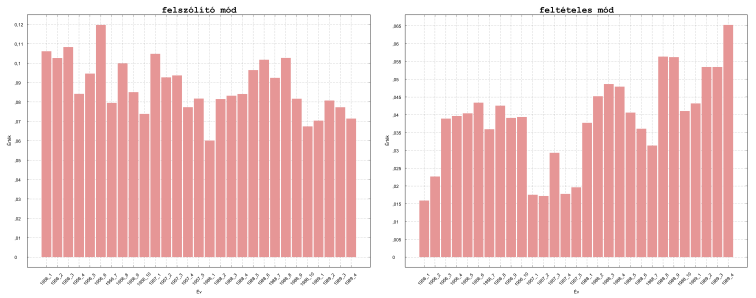
4.1. Morfológiai eltérések

A különböző igemódok használatát illetően azt tapasztaltuk, hogy az első periódusban jóval gyakrabban szerepelnek a felszólító módú igék. Ebből arra következtethetünk, hogy a diskurzus egyes cselekvések elvégzésének szükségességét hangsúlyozhatja az állampárt akaratával összhangban. A fentebbivel szoros

Jellemző	**p-érték**	**Jellemző**	**p-érték**
mondatszám	0,0047	alanyok aránya	<0,0001
tokenszám	<0,0001	tárgyak aránya	<0,0001
mondathossz	<0,0001	jelzők aránya	0,0002
ismeretlen szavak száma	0,0431	határozók aránya	0,0062
ismeretlen szavak aránya	0,0326	mellérendelések aránya	0,0002
melléknevek száma	0,0171	episztemikus bizonytalanság száma	0,0005
tulajdonnevek száma	0,0465	feltételes bizonytalanság száma	0,0454
melléknevek aránya	0,0017	doxasztikus bizonytalanság száma	0,0003
számnevek aránya	0,0236	episztemikus bizonytalanság aránya	0,0001
felsőfok száma	0,0106	weasel aránya	0,0466
középfok száma	0,0003	peacock aránya	0,0109
felsőfok aránya	<0,0001	hedge aránya	0,0130
középfok aránya	0,0003	doxasztikus bizonytalanság aránya	<0,0001
E/1. igék száma	0,0016	félelem szavainak száma	0,0263
T/1. igék száma	0,0055	düh szavainak száma	0,0151
múlt idő aránya	<0,0001	szorongás szavainak száma	0,0001
jelen idő aránya	0,0029	öröm szavainak aránya	0,0167
feltételes mód száma	0,0007	szorongás szavainak aránya	<0,0001
felszólító mód aránya	0,0258	emotív negatív szavak száma	0,0073
feltételes mód aránya	0,0002	negatív szavak aránya	<0,0001
E/1. igék aránya	0,0005	emotív negatív szavak aránya	0,0010
T/1. igék aránya	0,0013	befelé forduló igék száma	0,0105
ható igék száma	0,0003	meggyőzés igéinek száma	0,0017
műveltető igék száma	0,0471	befelé forduló igék aránya	<0,0001
ható igék aránya	<0,0001	meggyőzés igéinek aránya	0,0001
műveltető igék aránya	0,047	idézetek aránya	0,0037

2. táblázat. Statisztikailag szignifikáns nyelvi jellemzők.

összefüggésben, ugyanebben az alkorpuszban a feltételes módú igék előfordulása jelentősen ritkább, míg a második periódust kifejezetten jellemzi a feltételes módú igék használata, amelyet a bizonytalanság jeleként értelmezhetünk (vö. az 1. ábra).



1. ábra: A felszólító és feltételes módú igék időbeli gyakorisági megoszlása.

A felsőfokú melléknevek és határozószók gyakori előfordulása az első periódusban egyfajta propagandatechnikának is tekinthető: használatuk önbizalmat fejez ki, segítségükkel kiemelhető az adott politikai ideológia jelentősége, valamint a rendszer megkérdőjelezhetetlensége.

Eredményeink korrelálnak egy korábbi dolgozattal (Propaganda Analysis, 1938), amely a propagandaszövegekben fellelhető nyelvi eszközöket vizsgálta. Ilyen például a „glittering generality" jelensége, melynek értelmében a kiemelkedően nagyra becsült fogalmak és hiedelmek a hozzájuk kapcsolódó jelenségek általános és indoklás nélküli elfogadását váltják ki, említésükkel tehát mintegy manipulálható az olvasói elkötelezettség a szöveg tartalma iránt. Az úgynevezett „cherry picking"-elv alapján pedig a bizonyítékok elfedése, vagy a hiányos bizonyítékokon alapuló téves következtetések bizonyos adatokat és eseteket igazolnak, míg az ezeknek ellentmondó adatokat ignorálják.

Az igeidők vizsgálata a két alkorpuszban ugyancsak szignifikáns különbséget mutat. Így például az 1956-57-es periódusban jelentősen gyakoribbak a múlt idejű igék. Az adatok kézi elemzése megmutatta, hogy a múlt idejű igékkel a szövegek az 1956-os forradalom előzményeire, és az ahhoz vezető politikai döntésekre utalnak, az azokhoz vezető okokat, problémákat taglalják.

Megvizsgáltuk a két alkorpuszban előforduló igealakok személyét és számát is. Itt azt tapasztaltuk, hogy a második alkorpuszban lényegesen magasabb számban fordulnak elő a többes szám első személyű alakok, mint az elsőben. A két időszak adatainak összevető, kézi elemzése megmutatta, hogy amíg az 1956–57-es szövegek nagyrészt más, a pártvezetésen kívüli politikai és társadalmi szereplőkkel foglalkoznak (például azokkal a csoportokkal, amelyek cselekedetei és döntései a forradalomhoz vezettek), az 1988–89 a szövegek szerzői (akik maguk is a párt vezető tisztségviselői) javarészt saját magukról írnak.

4.2. Szemantikai eltérések

A szentiment- és emócióelemzés eredményeit illetően, a korszak elejének alkorpuszát kifejezetten jellemzik a pozitív érzelmeket kifejező elemek, mint például az öröm és a szeretet. Mindez összhangban van Propaganda Analysis (1938)-nek a már korábban említett eredményeivel: a propagandaszövegekben gyakoriak a pozitív érzelmeket kiváltó fogalmak és hiedelmek (mint például a *szerelem*, az *ország* és az *otthon*, valamint a *béke*, a *szabadság*, a *dicsőség* és a *becsület*) (vö. „glittering generality"). Ezzel szemben a negatív érzelmek, mint a félelem és a szorongás, gyakrabban fordulnak elő a 1988-89-es időszakban keletkezett szövegekben, amelyek jelzik a kommunista rezsim helyzetével, a közelgő változásokkal kapcsolatos félelmeket.

2. ábra: Különböző érzelmek időbeli gyakorisági megoszlása.

A fentiek mellett azonosítottuk a bizonytalanság nyelvi elemeinek különböző típusait is. Vizsgálataink azt mutatják, hogy háromféle bizonytalanság, a weasel (bizonytalan forrás), a peacock (túlzó kifejezések) és a hedge (mennyiségre vonatkozó bizonytalanság) sokkal gyakrabban jelenik meg az első alkorpuszban, mint a második időszakban (Vincze, 2013). A szemantikai bizonytalansággal szemben (Szarvas és mtsai, 2012) a diskurzusszintű bizonytalanság esetén „a hiányzó vagy szándékosan kihagyott információ nem a mondat propozíciós tartalmához kapcsolódik, hanem más tényezőkhöz", pl. a *néhány, gyakran, sok* stb. kifejezések nem adják meg a szóban forgó dolog vagy esemény pontos számát, illetve gyakoriságát. Eredményeink alapján a fentebbiekhez hasonló kifejezések okozta torzí-

tás a propagandadiskurzus jellegzetes vonásának tekinthető. A diskurzusszintű bizonytalansággal szemben az úgynevezett szemantikai bizonytalanság elemei (pl. *lehet, lehetséges, hisz* stb.) csupán ritkán fordulnak elő az első időszakban. Szemantikai bizonytalanság esetében a bizonytalanságjelölő szemantikai tartalma felelős a bizonytalanságért (Vincze, 2014). Például, az első alkorpuszban a szemantikai bizonytalanság episztemikus és doxasztikus (azaz hiedelmekre vonatkozó) típusai elég ritkának mondhatóak. Eredményeink ismét korrelálnak a propagandaszövegek nyelvi eszközeivel kapcsolatos korábbi kutatási eredményekkel (Propaganda Analysis, 1938): a „glittering generality" elvével összefüggésben, a propagandaszövegek egyik fontos jellemzője a pozitív érzelmeket kiváltó kifejezések gyakori használata, amelyek inherensen hordozzák a meggyőződést és nem igényelnek indoklást. Ezekből az eredményekből arra következtethetünk, hogy az első alkorpusz szövegei implicit módon megtévesztőbbek a második alkorpusz szövegeinél.

Az igéket további, szemantikai vizsgálatoknak is alávetettük a következőképpen: Megvizsgáltuk a „kifelé fordulás" igéit (public verbs, pl. *bejelent, fenntart, igazol* stb.), a „befelé fordulás "igéit (private verbs, pl. *reménykedik, ítél, feltételez* stb.), valamint az érvelés igéit (suasive verbs, pl. *egyetért, javasol, elismer* stb.). Az eredmények közül a legszembeötlőbb az volt, hogy a második vizsgált korszakban jelentősen gyakoribbak a befelé fordulás igéi a kifelé fordulás igéivel szemben, ami utalhat a rendszer elbizonytalanodására, a hatalom erejének a csökkenésére. Az érvelés igéi ugyancsak a második korszakban gyakoribbak, amire egy lehetséges magyarázat, hogy a gyengülő rendszerben a *Pártélet* szerzői fokozott figyelmet fordítottak a párt elveinek propagálására.

Az alkorpuszok szemantikai tartalmára vonatkozó vizsgálati eredményeink alapján összességében elmondható, hogy amíg az első időszak szövegei az erős és magabiztos, ugyanakkor a megtévesztés nyelvi jegyeit erősen magán hordozó kommunikáció jegyeit viselik magukon, a politikai korszak végét egy jelentősen kevesebb erőt sugárzó, félelemmel és idegességgel teli, önreflexív diskurzus jellemzi.

4.3. Lexikális eltérések

Végezetül összehasonlítottuk a két időszakban a különböző főnevek, igék és melléknevek gyakorisági megoszlását. Általánosságban elmondható, hogy alapvetően ugyanazok az elemek fordulnak elő mindkét alkorpuszban az egyes szófajok esetében, ugyanakkor az elemek gyakorisága jelentősen eltér. Az előbbi sajátság egyrészt a szövegek tematikájával, másrészt a propagandajelleggel is magyarázható. Az utóbbi tulajdonság ugyanakkor arra mutat, hogy a hangsúlyok jelentősen mások a politikai korszak végén annak elejéhez képest.

Az első alkorpuszban a legtöbbször előforduló főnevek között sok olyat találunk, amelyek szemantikailag a vezetőség hatalmával, rendelkezéseivel kapcsolatosak (pl. *terv, teljesítés* stb.). Ugyanakkor a második alkorpuszban a leggyakrabban előforduló főnevek között sok, az eltérő vélemények és a választás lehetőségére utaló elem szerepel (pl. *lehetőség, döntés, vélemény* stb.).

A főnevek mellett a melléknevek gyakorisági megoszlása is jól érzékelteti a két korszak közötti különbséget. Az eltéréseket a 3. ábra szófelhőivel szemléltetjük.

3. ábra: A melléknevek gyakorisága a két vizsgált korszakban (bal oldal: 1956-57, jobb oldal: 1988-89).

Azt látjuk, hogy amíg például az első vizsgált időszak diskurzusában a *szocialista* és a *dolgozó* fogalmak különösen magas frekvenciával szerepelnek, addig ezek az elemek a politikai időszak végére jelentősen háttérbe szorulnak. Ugyancsak figyelemre méltó sajátság, hogy az első alkorpuszban a *gazdasági* fogalomnak kiemelkedő szerep jut. Ez bizonyosan összefüggésben áll azokkal a gazdasági reformokkal, amelyek a Kádár-korszak elejének politikájában kulcsszerepet töltöttek be (vö. a mezőgazdaság erőszakos kollektivizálása). Ugyanakkor a korszak második felétől ezek a reformok jelentősen háttérbe szorultak, amelyet tükröznek a második alkorpusz melléknévi gyakorisági adatai is. A *gazdasági* például átadja kiemelt szerepét a *társadalmi* fogalomnak, amely korábban nem bírt nagy jelentőséggel.

5. Konklúzió

Dolgozatunkban a Kádár-korszak elejének és végének politikai diskurzusának elemzését végeztük el a *Pártélet* című pártlap szövegeinek nyelvi elemzésén keresztül. Amint a dolgozatban részletesen ismertettük, a *Pártélet* mint az országot irányító állampárt hivatalos kiadványa a politikai ideológia terjesztését célozta. Ezzel összefüggésben a lap a propaganda mint diskurzus vizsgálatának kiemelt fontosságú anyagának tekinthető. Dolgozatunkban két, megközelítőleg egyforma hosszúságú időszakot vizsgáltunk: a 1957 januárja és 1958 decembere közötti, valamint a 1988 januárja és 1989 áprilisa közötti időszakot.

A vizsgálat során a szövegek különböző morfológiai, szintaktikai és szemantikai jellemzőit elemeztük azok gyakorisági megoszlására vonatkozóan, és nem csupán a két korszak, de az egyes lapszámok közötti eltéréseket is kiszámítottuk és vizsgáltuk. Vizsgálataink több tekintetben is szignifikáns eltérést mutattak a két alkorpusz között. Eredményeink pedig korrelálnak a propagandaszövegek nyelvi eszközeivel kapcsolatos korábbi kutatási eredményekkel.

A dolgozatban bemutatott eredményekkel a politikai propaganda diskurzusának kutatásához kívántunk hozzájárulást tenni, különös tekintettel a kommunis-

ta korszak propagandanyelvének elemzésére. Annak céljából, hogy megállapításainkra alapozva a vizsgált diskurzusról általános érvényű jellemzést adhassunk, a kutatás fontos, további lépése volna, hogy az itt alkalmazott elemzési módszerek és eszközök segítségével a vizsgált korszakban keletkezett, nem propagandisztikus szövegek elemzését is elvégezzük, eredményeinket azok nyelvi sajátságaival összehasonlítsuk. Itt azonban fontos megjegyeznünk, hogy a kommunista korszakból – a párt direkt politikai agitációjával összefüggésben – nem áll rendelkezésünkre olyan kiadvány, amelynek tartalmát bizonyosan semleges, propagandától „mentes" szövegnek tekinthetnénk. A fentebbi összevető vizsgálatot tehát nem tartjuk lehetségesnek. E problémával összefüggésben azt tervezzük, hogy a kutatás következő lépéseként rendszerváltást követően publikált sajtóanyagokat vizsgálunk, és azokat vetjük össze az itt bemutatott eredményeinkkel.

Munkánk hosszú távú célja egy olyan elemző algoritmus létrehozása, amely képes hatékonyan elkülöníteni a politikai agitációt más, propagandától mentes szövegektől.

Köszönetnyilvánítás

A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM).

Hivatkozások

Barrón-Cedeño, A., Jaradat, I., Martino, G.D.S., Nakov, P.: Proppy: Organizing the news based on their propagandistic content. Information Processing Management 56(5), 1849–1864 (2019), https://doi.org/10.1016/j.ipm.2019.03.005

Dér, Cs.I., Markó, A.: A magyar diskurzusjelölők szupraszegmentális jelöltsége. In: Nyelvelmélet–nyelvhasználat, pp. 61–67. Tinta, Székesfehérvár–Budapest (2007)

Fetzer, A.: „And I Think That Is a Very Straightforward Way of Dealing With It"– The Communicative Function of Cognitive Verbs in Political Discourse. Journal of Language and Social Psychology 27, 384–396 (12 2008)

Fraser, B.: Chapter 11. hedging in political discourse. In: Perspectives in Politics and Discourse, pp. 201–214 (01 2010)

Girlea, C., Girju, R., Amir, E.: Psycholinguistic Features for Deceptive Role Detection in Werewolf. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 417–422. Association for Computational Linguistics, San Diego, California (Jun 2016)

Graves, L.: Understanding the promise and limits of automated fact-checking. Tech. rep., Reuters Institute – University of Oxford (2018)

Greene, S., Resnik, P.: More than words: Syntactic packaging and implicit sentiment. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 503–511. Association for Computational Linguistics, Boulder, Colorado (Jun 2009), https://www.aclweb.org/anthology/N09-1057

Jalilifar, A.R., Alavi, M.: Power and Politics of Language Use: A Survey of Hedging Devices in Political Interviews. The Journal of Teaching Language Skills (JTLS) 3, 43–66 (2011)

Propaganda Analysis, I.f.: How to Detect Propaganda. In: Propaganda Analysis. Publications of the Institute for Propaganda Analysis. vol. I, pp. 210–218 (1938)

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of the English Language. Longman, London (1985)

Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2931–2937. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)

Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: Three types of fakes. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community. pp. 83:1–83:4. ASIST '15, American Society for Information Science, Silver Springs, MD, USA (2015), http://dl.acm.org/citation.cfm?id=2857070.2857153

Scheithauer, R.: Metaphors in election night television coverage in Britain, the United States and Germany. In: Political Discourse in the Media: Cross-cultural perspectives, pp. 75–106 (01 2007)

Simon-Vandenbergen, A.M., White, P., Aijmer, K.: Presupposition and 'taking-for-granted' in mass communicated political argument An illustration from British, Flemish and Swedish political colloquy. In: Political Discourse in the Media, pp. 31–74 (01 2007)

Szabó, M.K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In: Segédkönyvek a nyelvészet tanulmányozásához 177, pp. 278–285. Tinta, Budapest (2015)

Szabó, M.K., Ring, O., Nagy, B., Kiss, L., Koltai, J., Berend, G., Vidács, L., Kmetty, Z.: Mapping the dynamic change of the concept „industry" and „agriculture" in the Hungarian Socialist era using a word embedding model. Historical Methods: A Journal of Quantitative and Interdisciplinary History. Publikálásra benyújtva (2019)

Szabó, M.K., Vincze, V., Morvay, G.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: Távlatok a mai magyar alkalmazott nyelvészetben, pp. 282–292. Tinta, Budapest (2016)

Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. Computational Linguistics 38, 335–367 (June 2012)

Thorne, J., Vlachos, A.: Automated fact checking: Task formulations, methods and future directions. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3346–3359. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1283

Vincze, V.: Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In: Proceedings of the Sixth International Joint Conference

on Natural Language Processing, Nagoya, Japan, October 2013. pp. 383–391 (2013)

Vincze, V.: Uncertainty Detection in Natural Language Texts. Ph.D.-értekezés, University of Szeged, Szeged, Hungary (2014)

Wang, W.Y.: "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426. Association for Computational Linguistics, Vancouver, Canada (July 2017), http://aclweb.org/anthology/P17-2067

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)

Német-magyar nyelvtanulói korpusz (Dulko)

Kappel Péter[1], Modrián-Horváth Bernadett[1],
Andreas Nolda[2], Vargáné Drewnowska Ewa[1]

[1] Szegedi Tudományegyetem, Germán Filológiai Intézet
kappelp@lit.u-szeged.hu,
{bernadett.modrianhorvath,ewa5drewa}@gmail.com
[2] Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23,
10117 Berlin, Germany
andreas@nolda.org

Kivonat: Cikkünkben bemutatjuk a Dulko korpuszt, amely magyar anyanyelvű, a németet mint idegen nyelvet tanuló egyetemisták által létrehozott szövegeket tartalmaz. A német-magyar nyelvtanulói korpusz várhatóan 2020-tól szabadon hozzáférhető és (több nyelvtanulói korpuszhoz hasonlóan) az ANNIS keresőrendszerrel (Krause és Zeldes, 2016) online kutatható lesz. A nyelvtanulói szövegek többszintű annotációja a szóalakokon kívül többek között lemmán, szófajokon, metaadatokon (pl. a célnyelv tanulásával töltött időtartam) és hibakategóriákon alapuló lekérdezéseket is lehetővé tesz. A korpusz építése során mind tartalmi, mind korpusztechnológiai szempontból számos innovatív elemet alkalmazunk. Kiemelendő egyrészt az explicit hibajelölés és -kategorizálás, másrészt egy olyan nyílt forráskódú szoftver kifejlesztése, amely (többek között beépített lemmatizálóval és szófaji egyértelműsítővel) megkönnyíti a német nyelvű szövegek annotációját, és a nyelvtanulói korpuszokkal szemben támasztott elvárásokhoz igazodva lehetővé teszi a nyelvi adatok többszintű, transzparens elemzését.

Kulcsszavak: nyelvtanulói korpusz, korpuszépítés, hibaannotáció

1 DULKO - egy új német-magyar nyelvtanulói korpusz szükségességéről

A Dulko korpusz (*D*eutsch-*u*ngarisches *L*erner*ko*rpus) része egy nemzetközi projektnek, amelyben három terület kap központi szerepet: a korpusztechnológia, a nyelvészet és a nyelvdidaktika.[1] A Dulkót a projekt futamideje, azaz három év alatt, a Szegedi Tudományegyetemen tanuló germanisztika szakos hallgatók által írt esszék és fordítások alapján hozzuk létre és tesszük szabadon hozzáférhetővé (CLARIN PUB+BY+SA+PRIV). A Dulko projekt elsődleges célját a germanista hallgatók írásbeli szövegalkotásában jelentkező nyelvi hibák[2] empirikus vizsgálata képezi. A kor-

[1] A részletekhez az említett nemzetközi projektről vö. http://arts.u-szeged.hu/kutatas-tudomany/dulko

[2] Itt fontosnak tartjuk előre leszögezni, hogy germanista hallgatók által vétett nyelvi hibákat eltéréseknek tekintjük a nyelvtanuló köztes nyelvének rendszere és a célnyelv rendszere kö-

puszba a nyelvi hibákkal kapcsolatos adatokat, pl. az egyes hibatípusokat is integráljuk, ezzel a tipikus lexikai, nyelvtani és helyesírási hibák elektronikusan is kereshetővé válnak.

Ezzel a célkitűzéssel a Dulko a nyelvelsajátítás-kutatást és korpusznyelvészetet kapcsolja össze. Korpusznyelvészeti szempontból nézve feltehető a kérdés, hogyan lehet nyelvi eltéréseket egy tanulói korpuszban jól áttekinthetően és nyomon követhető módon a célhipotézisek felállítása által (vö. Reznicek és mtsai, 2013) hibaként interpretálni, kategorizálni és többdimenziós szófaji annotáció és lemmatizáció segítségével a mai elvárások szerint kereshetővé tenni. A projekt céljának megvalósítására csoportunk egyik munkatársa egy olyan eljárást dolgozott ki, amely lehetővé teszi a hallgatók szövegeinek elektronikus feldolgozását nyelvtanulói korpusz formájában (vö. 2-4. fejezet).

Ezen korpusz létrehozása mellett három egymással összefüggő érv szól. Mindenekelőtt figyelemre méltó a korpusz sajátossága és egyedülállósága nyelvészeti és nyelvdidaktikai szempontból. A Dulko szövegei szerzőinek anyanyelve és az általuk megfogalmazott szövegek nyelve nyelvtipológiailag lényegesen különböznek egymástól, ebben rejlik a Dulko korpusz sajátossága pl. a rokon nyelveken alapuló tanulói korpuszokkal szemben. Az indogermán német és a finnugor magyar nyelv nem állnak genetikai rokonságban egymással. Morfoszintaktikai szempontból nézve a német egy leginkább fuzionáló-analitikus-izoláló nyelv, míg a magyar elsősorban agglutináló nyelvnek számít. Ebből számos különbség következik a két nyelv között a fonetikai, fonológiai, morfológiai, szintaktikai, lexikológiai, frazeológiai, pragmatikai és szövegnyelvészeti szinten.[3] Ezek a nyelvtipológián alapuló kontrasztok sok potenciális hibaforrást[4] képezhetnek a magyar anyanyelvű németül tanulók számára. Számos hiba vezethető vissza a német és a magyar mondatok eltérő információs szerkezeti felépítésére vagy olyan különbségekre, mint a nyelvtani nem megléte a németben és hiánya a magyarban (lásd pl. 1. ábra). A Dulko egyik központi tartalmi célkitűzése az annotált nyelvi hibák leírása és visszavezetése a két nyelv között fennálló nyelvtipológiai különbségekre.[5] A korpusz másik ehhez kapcsolódó célja, hogy megfelelő adatokat szolgáltasson a tanulói nyelv (Lernersprache) kutatásához.[6] A tanulói nyelv egy még mindig kevéssé feltárt, nagyon releváns kutatási területnek számít. Hazai és nemzetközi viszonylatban a legtöbb publikáció a kezdő szintű tanulói nyelvváltozat kutatásával foglalkozik, a német-magyar nyelvtanulói korpuszunk sajátossága tehát

zött. A 'köztes nyelv' (másként: 'interimnyelv') – a nyelvtanulás folyamata során kialakult sajátos nyelvrendszer, amely úgy a tanuló anyanyelve, mint az általa elsajátítandó a célnyelv jellemzőit tartalmazza. Ezen kívül a köztes nyelv rendszerében más jellemzők is találhatók, amelyek sem az anyanyelvben, sem a célnyelvben nem lépnek fel. (vö. Selinker, 1972; Fekete 2016).

[3] Részletes nyelvtipológiai leírásokhoz a német és a magyar nyelv között vö. Brdar-Szabó, 2010b; Gunkel és mtsai, 2017; Pilarský, 2018.

[4] A cikk korlátozott terjedelme miatt itt nem térünk ki más releváns faktorokra, amelyek szintén hibákhoz vezethetnek a német nyelvű szövegek fogalmazásánál.

[5] Ezzel együtt e célkitűzés a német és a magyar nyelv közötti esetleges hasonlóságokat természetesen nem hagyja figyelmen kívül. A fő elv itt az, hogy a hibaelemzés a célnyelv és a tanulók anyanyelvének összevetésén alapuljon.

[6] A nyelvi hiba, a hibaelemzés és a nyelvek közötti kontraszt fontosságához a nyelvelsajátításban és nyelvtanításban vö. Brdar-Szabó, 2010a; Fekete, 2008, 2016.

többek között abban rejlik, hogy kontrollált körülmények között, haladó nyelvtanulóktól gyűjtött autentikus adatokból áll.[7]

Második érvként az olyan tananyagok hiánya jelölhető meg a német mint idegen nyelv tanításában, melyek a következő, a jelenlegi Nemzeti Alaptantervben (2018) megnevezett tanulói kompetenciák fejlesztését segítenék: „Az anyanyelv és az idegen nyelv különbségének felismerése, ennek megfogalmazása a diák saját szavaival"; „Az anyanyelvi és az idegen nyelvi ismeretek összevetése, az egyes jelenségek egyre pontosabb megnevezése"; „Az anyanyelvhez és az idegen nyelvhez kötődő sajátosságok összevetése az általános nyelvészeti ismeretek felhasználásával."[8] Ezeknek az elvárt kompetenciáknak a magyar iskolákban a német nyelv tanításához használt tankönyvek alig felelnek meg, mivel egyáltalán nem, vagy csak nagyon ritkán kerül bennük szóba a német és a magyar nyelv összevetése. Erről meggyőződhettünk, amikor 2016-ban a projektünk elindítása előtt egy felmérést végeztünk a szegedi gimnáziumi tanárok között. Hasonló eredményekre jutott Fekete (2016), aki a „Schulbus", „Das Deutschmobil", „Start! Neu" és „Unterwegs" c. tankönyveket abból a szempontból vizsgálta meg, hogy a nyelvtan ismertetése mennyire alapul a nyelvi különbségek figyelembe vételén. A német és magyar nyelv összevetésének teljes hiánya, vagy a nyelvi különbségek csak nagyon következetlen figyelembevétele ezekben a tankönyvekben nyomós okot szolgáltat arra, hogy kiegészítő, rendszeresen használható és progresszíven felépített tananyag készüljön. Itt a Dulko korpusz potenciális forrásként és alapként szolgálhat az új tananyagok kialakításához.

A Dulko tudomásunk szerint az egyetlen, az ANNIS keresőrendszerrel (Krause és Zeldes, 2016) online kutatható német-magyar nyelvtanulói korpusz (a részletekhez l. 4. és 5. fejezet). Ennek köszönhetően az érdeklődő szakemberek számára könnyen és széles körben hozzáférhetővé válik. Itt elsősorban német nyelvtanárokra és nyelvi kontraszttal, nyelvtipológiával foglalkozó nyelvészekre gondolunk.[9]

Harmadik, záró érvként megemlítendő, hogy a Dulkónak a korpusztechnológia területén belül is fontos szerep jut: A tanulmány elején megnevezett nemzetközi projektnek, amelynek a Dulko részét képezi, többek között az a célja, hogy olyan eljárásokat dolgozzunk ki, melyek lehetővé teszik, hogy nyelvi tulajdonságokat szövegkorpuszok alapján egy elemzést támogató szoftver segítségével hasonlíthassunk össze. Ennek előfeltétele az összehasonlítható német és magyar nyelvű korpuszok fejlesztése. Az összehasonlítható korpuszok kialakításánál a DeReKo (Deutsches Referenzkorpus, a német nyelv reprezentatív korpusza, vö. Institut für Deutsche Sprache, 2004ff.) és az MNSZ (Magyar Nemzeti Szövegtár, vö. Váradi, 2002) korpuszokra támaszkodunk. A két korpusz technológiai harmonizációját a mannheimi IDS által

[7] Juhász (1970) az interferencia problémáival foglalkozó monográfiája nem tanulói korpuszon, hanem kísérleti alapon gyűjtött, kevésbé autentikus adatokon alapul. A Falko alkorpuszokban csak nagyon kevés adat van magyar anyanyelvű tanulókról (vö. Reznicek és mtsai, 2012). Fekete (2016) longitudinális elemzése egy 90 írott szövegből álló korpuszon alapul. A szövegek magyar gimnazistáktól származnak. Ezzel szemben a Dulko korpusz esetében haladó tanulói nyelvváltozatról van szó Walter és Grommes (2008) értelmében. Ezen kívül Fekete korpusza online nem elérhető.

[8] Vö. NAT (2018) „Anyanyelvi kultúra, ismeretek az anyanyelvről" c. 5. fejezete 37. o.

[9] Az ANNIS-keresést támogató formátum segítségével a Dulkóban levő annotátumok integrálhatók lesznek a Falko korpuszba (a részletes technikai leíráshoz vö. 3.1 fejezet). Ez is lényegesen növelheti a Dulko korpusz nyilvános jellegét és hozzáférhetőségét.

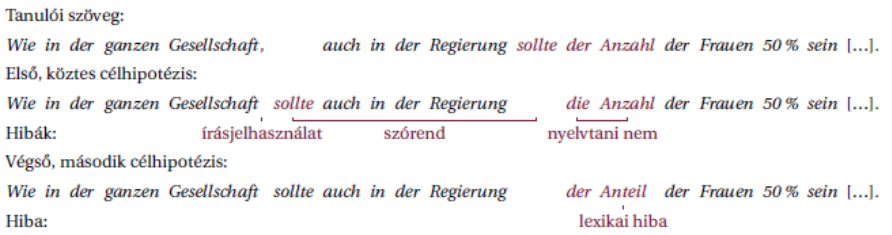
fejlesztett KorAP rendszerrel (Korpusanalyseplattform der nächsten Generation, a következő generáció korpuszelemzési plattformja)[10] kívánjuk megoldani. A Dulko korpuszt nyelvi hibák annotációjának továbbfejlesztése mellett az újgenerációs KorAP rendszerébe is be szeretnénk ágyazni.

2 Az annotációs eljárás alapelvei

A Dulko korpusz adatgyűjtés és -kezelés tekintetében a Falko korpuszon alapszik. A Falko egy nyelvtanulói szövegkorpusz, amelyet 2005 óta a berlini Humboldt Egyetemen fejlesztenek (vö. Reznicek és mtsai, 2012). A Falkótól azonban abban különbözik a Dulko, hogy itt a nyelvi hibák explicit többdimenziós (többszintes) annotációját is elvégezzük. Így a Dulko a következő releváns pontokban tér el a Falkótól (vö. Hirschmann és Nolda, 2019; Nolda, 2019):

1. A Dulko-féle annotációs eljárásban a célhipotézisek mennyisége tetszés szerint adható meg.
2. A Dulko-eljárásban a hibák és ezek területei explicit módon annotálhatók a hibakategóriák segítségével különböző nyelvi szinteken.
3. A Dulko-eljárásban minden célhipotézishez hozzá lehet rendelni hibakategóriát bármely nyelvi szinten.

A célhipotézisek segítségével lépésenként közelíthető a tanulói szöveg a hibáktól megtisztított célnyelvi megfelelőjéhez. Az alábbi szövegrészletben a következő eltéréseket tekinthetjük hibának (vö. 1. ábra):



1. ábra: Egy esszészöveg részletének hibaelemzése, 2017/2018. I. félév, SZTE

- A tanulói szövegben kitett vessző helyesírási hibának számít.
- Szórendi hiba van a *sollte* igealak esetében.
- A *der Anzahl* főnévi csoport *der* névelőjén jelzett nyelvtani neme hibás.
- A *der Anzahl* főnévi csoport *Anzahl* lexikai egysége nem illik a kontextushoz, lexikai hibaként értékelhető: az *50 %* nem számot (*Anzahl*), hanem arányt (*Anteil*) fejez ki.

[10] http://www1.ids-mannheim.de/kl/projekte/korap/

Amennyiben nem adnánk meg az első, köztes célhipotézist, a *der Anzahl* hibás nyelvtani német nem lehetne explicit formában láthatóvá tenni, mivel az *Anzahl* főnév nyelvtani neme megegyezik az *Anteil* főnévével.

Ahogy a fenti példa mutatja, a köztes célhipotézisek azokat a hibákat teszik láthatóvá, amelyek a végső hipotézisnél az átfedések miatt láthatatlanná válnak. A köztes célhipotézishez igény szerint alternatív hipotézisek is megadhatók (vö. Nolda, 2019).

Köztes célhipotézisekre akkor is szükség lehet, amikor például morfológiai vagy lexikai és szórendi hiba együtt fordul elő, ugyanazon a nyelvi szinten két vagy több hiba található, illetve ha más okból nem oldható meg egy lépésben a korrektúra.

3 Az EXMARaLDA (Dulko) annotációs szoftver

Projektünk koncepciója a Falko korpuszcsalád nyomán, arra nagymértékben építve alakult ki, így annotációs eljárásunk kiindulópontja is a Falkóban használt konvenció volt. Ezt Andreas Nolda munkatársunk Hagen Hirschmann közreműködésével továbbfejlesztette és az EXMARaLDA-Partitur-Editor programra[11] implementálta. 2018-ban az „Innovatív eljárás idegennyelv-tanulók nyelvi adatainak annotációjára nyelvtanulói korpuszokban: Koncepció, modellálás, programozás” munkájáért a Szegedi Tudományegyetem innovációs díját kapta a műszaki tudományok területén.

A Dulko (akárcsak a Falko) annotációs eljárása automatikus és manuális elemeket is magába foglal. Automatikus elsősorban a tokenizálás, a mondathatárok szerinti, a szófaj szerinti (pos) annotáció, a lemmatizáció, valamint a tanulói szöveg és a célhipotézis, ill. az egyes kumulatív célhipotézisek közti eltérések annotációja. A manuális annotáció főként a tanulói szöveg egyes mondataihoz rendelt célhipotézisekre, valamint az eltérések különböző nyelvi szinteken való explicit hibaannotációjára irányul.

Az alábbiakban röviden bemutatjuk az annotációs program működési mechanizmusát, valamint a főbb transzformációs műveleteket.

3.1 Az annotációs program működési háttere: felhasznált szoftverek és kereshetőség

A program az annotációs lépéseket az EXMARaLDA-Partitur-Editor (vö. Schmidt, 2004) segítségével végzi, ennek módosításaként jött létre a tanulói adatokat annotáló EXMARaLDA (Dulko). Ez egy nyílt forráskódú szoftver, mely a Bitbucket projekt hosting platformon keresztül ingyenesen hozzáférhető (a licenc GPL, 2. verziójú).[12] Hasonló programról nincsen tudomásunk. A program futtatható Linux, MacOS és Windows operációs rendszerekkel is.

Az EXMARaLDA (Dulko) projektspecifikus transzformációs szcenáriókat tartalmaz, amelyek mindegyike egy, a share-jegyzékben tárolt XSLT-stylesheethez kapcsolódik. A tokenizálás után, melyet az EXMARaLDA-Partitur-Editor végez, történik a mondathatárok szerinti annotáció, a pos-tagging és a lemmatizálás a TreeTagger (Schmid, 1997) programon belüli alkalmazásával. A TreeTagger a német nyelv szófa-

[11] https://exmaralda.org/de

[12] https://bitbucket.org/nolda/exmaralda-dulko/downloads/

ji meghatározásánál sztenderdnek számító STTS-tagsetet (Schiller és mtsai, 1997) használja. Ezt követi a célhipotézisek, ill. a célhipotézisek és a tanulói szöveg közötti eltérések regisztrálása. Az EXMARaLDA (Dulko) projektspecifikus transzformációi közé tartozik t. k. a tanulói produktum célhipotézisbe másolása, ill. – kumulatív célhipotézisek esetén – a célhipotézisek következő szintű célhipotézisbe való másolása, a hibaannotálás célhipotézisenként négy szintjének kialakítása, valamint ide kapcsolódik a hibatageket tartalmazó XML-dokumentum, az annotációs panel is.

Az annotáció során nyert XML-dokumentumot egy további EXMARaLDA (Dulko)-transzformáció segítségével az ANNIS keresővel (Krause és Zeldes, 2016) kompatibilis formátumba lehet hozni, illetve html-formátumba is lehet konvertálni. Az előbbi az internetes (vagy helyi hálózatokon belüli) ANNIS-keresést[13] támogatja, ezáltal az annotátumok integrálhatók lesznek a Falko korpuszba. A html-verzió egy mondatonként tördelt változatot tartalmaz, amely közvetlen olvasásra a leginkább alkalmas. Céljaink között szerepel a mannheimi Institut für Deutsche Sprache KorAP rendszerével való kompatibilitás kialakítása is (vö. 1.).

3.2 Transzformációk az EXMARaLDA (Dulko) szoftverben

A projektspecifikus transzformációk elsősorban a következőket tartalmazzák:

1. Metaadatok átvitele a Dulko-template-ből: az egész projektre érvényes adatok importálása, és a hallgatóra vonatkozó, valamint a szöveggel kapcsolatos metaadatok beviteléhez alkalmas sablonok átvétele. Az adatközlők anonimitásának megőrzése érdekében a hallgatói kódokból md5-kódokat generálunk és ezeket a program az annotáció minden sorához hozzárendeli.
2. Word-(szóalak-)sor (korábban: tok-sor) generálása: a bemásolt hallgatói szöveg tokenizálása, ill. ennek aktualizálása.
3. A hallgatói szöveghez (word-sor) kapcsolódó mondathatárok szerinti és pos-annotáció, valamint lemmatizáció és ezek aktualizálása.
4. Orig-sor, layout-sor és graph-sor: a program új (17.0-tól) verziójában lehetőség van a hallgatók által véghez vitt javítások (kihúzás, betoldás, javítás stb.) és az eredeti szöveg sortöréseinek és bekezdéseinek jelölésére. Erre a célra a word-sor adatai másolódnak az orig-, layout, ill. graph-sorokba, ezeket lehet manuálisan megváltoztatni.
5. Különbségek felismerése a word- és az orig-/layout-/graph-sorok között, valamint ezek aktualizálása.
6. Trans-sor (fordítás) hozzáadása: fordított szövegek esetén mondatonként kerül rögzítésre a fordítás alapját képező forrásnyelvi feladatszöveg.
7. Célhipotézis-sor és a hozzá kapcsolódó hibasorok létrehozása, ill. aktualizálása. A célhipotézisnél szintén a word-sorban található hallgatói szöveg kerül átmásolásra, melyet manuálisan lehet módosítani a célnyelvi normának megfelelően. Lehetőség van második, ill. további célhipotézisek hozzáadására is. A kapcsolódó hibasorok négy nyelvi szintnek megfelelő sorokat rendelnek minden célhipotézishez (helyes-

[13] https://corpus-tools.org/annis/

írás, morfológia, szintaxis és szemantika), melyekbe az annotációs panelben definiált hibakategóriák szerint lehet hibatageket beszúrni.
8. A célhipotézishez (vagy második, ill. további célhipotézishez) kapcsolódó mondathatárok szerinti és pos-annotáció, valamint lemmatizáció, valamint ezek aktualizálása.
9. Különbségek felismerése a word- és a célhipotézis-sorok között, ill. kumulatív célhipotézisek esetén az egyes célhipotézisek között (betoldás, törlés, mozgatás, egyesítés, hasítás) és ezek aktualizálása.
10. „Tisztító" transzformációk: a mondatszakaszok rendezésére, ill. a már nem használt időpontok[14] törlésére vonatkozó műveletek.
11. Konvertálás: html-, ill. ANNIS-kompatibilis verzió létrehozása.

3.3 Az annotációs eljárás szemléltetése

Az annotáció főbb lépéseinek szemléltetésére álljon itt néhány képernyőkép, melyeket egy fordításszöveg egy mondatának annotálása során készítettünk. Az elsőn (2. ábra) a tokenizált tanulói szöveg látszik, a következőn a mondathatár és szófaj szerint taggelt, lemmatizált változat.

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.

2. ábra Tokenizált hallgatói szöveg az EXMARaLDA (Dulko) programban

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1															
[pos]	PPER	VMFIN	PIS	VVINF	APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen	ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.

3. ábra Tokenizált, mondathatár és szófaj szerint annotált és lemmatizált hallgatói szöveg az EXMARaLDA (Dulko) programban

A 4. ábrán a program beszúrta a fordítási sort, melybe manuálisan bekerült a kiinduló nyelvi szöveg. Ezután a célhipotézis előállításához automatikusan bemásolásra kerül a word-sor (a tokenizált tanulói szöveg), valamint ekkor jelennek meg a hibataggelésre szolgáló sorok is (5. ábra). Ezeket manuálisan kell megváltoztatni, majd a különbségek felismerésére szolgáló transzformáció segítségével megjeleníteni a tanulói szöveg és a célhipotézis közötti eltéréseket (6. ábra).

Szükség esetén – például, ha egy nyelvi-annotációs szinten több hiba fordul elő, vagy egyidejűleg szórendi és más hiba is előfordul – második, illetve további célhipotézisek előállítására is sor kerülhet, melynek másolása, módosítása és annotálása az első célhipotézisével analóg módon történik; a végeredményt a 7. ábra mutatja. A két célhipotézist jelen esetben kumulatívan kell értelmezni, tehát a hallgatói szövegtől a lehető legkevésbé eltérő, de a célnyelvi normának már megfelelő változat itt a 2. célhipotézisben olvasható.

[14] Mivel az EXMARaLDA egy eredetileg beszélt nyelvi korpuszok annotációjára készült program, itt az egyes tokenek időpontokhoz vannak rendelve.

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1															
[pos]	PPER	VMFIN	PIS	VVINF	APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen	ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.															

4. ábra Fordítási sor hozzáfűzése az EXMARaLDA (Dulko) programban

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1															
[pos]	PPER	VMFIN	PIS	VVINF	APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen	ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.															
[ZH]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[FehlerOrth]																
[FehlerMorph]																
[FehlerSyn]																
[FehlerLex]																
[FehlerSem]																

5. ábra Célhipotézis automatikus előállítása az EXMARaLDA (Dulko) programban

4 A korpuszépítés folyamata

A tanulói korpusz építésében hét projekttagunk, valamint pályázat útján kiválasztott hallgatói segéderők vesznek részt. A korpusz építése a mintavétel megtervezésétől az adatgyűjtésen át az annotációs folyamat végrehajtásáig, illetve a korpusz publikálásáig tart. Az adatok feldolgozásával párhuzamosan az első időszakban a hibatagset optimalizálása is fontos feladatot jelentett projektcsoportunk számára, hiszen egyrészt néhány hibatípus csak nagyobb mennyiségű szöveg kiértékelése után fordult elő, másrészt bizonyos hibatagek összevonhatónak bizonyultak az annotációs munka során.

4.1 Adatgyűjtés, metaadatok gyűjtése

Az adatgyűjtés félévente történik intézetünkben kontrollált körülmények között, részben tanóra, részben vizsga keretében. A feladat (esszéírás vagy fordítás) elvégzéséhez a hallgatók semmilyen segédeszközt (szótárat, internetet stb.) nem vehetnek igénybe, a munka elkészítése – ellentétben a Falko szövegeinek legnagyobb részével – kézírással történik. A mintavétel megtervezése során törekszünk a longitudinális vizsgálatok lehetővé tételére, tehát ugyanazon hallgatói csoportoktól igyekszünk több egymást követő félévben is adatokat gyűjteni.

[word]	Ich	konnte	vieles	besuchen		ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1																
[pos]	PPER	VMFIN	PIS	VVINF		APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen		ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.																
[ZH]	Ich	konnte	vieles	besuchen	,	ohne	dass		ich	von	den	lokalen	Menschen	bemerkt	worden	wäre	.
[ZHDiff]					MOVT			MOVS		CHA							
[ZHS]	s1																
[ZHpos]	PPER	VMFIN	PIS	VVINF	$,	APPR	KOUS		PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[ZHlemma]	ich	können	viele	besuchen	,	ohne	dass		ich	von	die	lokal	Mensch	bemerken	werden	sein	.
[FehlerOrth]					ZS			ZS									
[FehlerMorph]																	
[FehlerSyn]										ValV							
[FehlerLex]																	
[FehlerSem]																	

6. ábra Manuálisan módosított, hibataggelt célhipotézis a hallgatói szövegtől való eltérések jelölésével az EXMARaLDA (Dulko) programban

[word]	Ich	konnte	vieles	besuchen		ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1																
[pos]	PPER	VMFIN	PIS	VVINF		APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen		ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.																
[ZH]	Ich	konnte	vieles	besuchen	,	ohne	dass		ich	von	den	lokalen	Menschen	bemerkt	worden	wäre	.
[ZHDiff]					MOVT			MOVS		CHA							
[ZHS]	s1																
[ZHpos]	PPER	VMFIN	PIS	VVINF	$,	APPR	KOUS		PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[ZHlemma]	ich	können	viele	besuchen	,	ohne	dass		ich	von	die	lokal	Mensch	bemerken	werden	sein	.
[FehlerOrth]					ZS			ZS									
[FehlerMorph]																	
[FehlerSyn]										ValV							
[FehlerLex]																	
[FehlerSem]																	
[ZH]	Ich	konnte	vieles	besuchen	,	ohne	dass		ich	von	den	Einheimischen		bemerkt	worden	wäre	.
[ZHDiff]												MERGE					
[ZHS]	s1																
[ZHpos]	PPER	VMFIN	PIS	VVINF	$,	APPR	KOUS		PPER	APPR	ART	NN		VVPP	VAPP	VAFIN	$.
[ZHlemma]	ich	können	viele	besuchen	,	ohne	dass		ich	von	die	Einheimische		bemerken	werden	sein	.
[FehlerOrth]																	
[FehlerMorph]																	
[FehlerSyn]																	
[FehlerLex]												Lex					

7. ábra Tokenizált hallgatói szöveg az EXMARaLDA (Dulko) programban

A szövegeken kívül egyúttal a hallgatókra vonatkozó metaadatok gyűjtésére is sor kerül, a Granger és Paquot (2017) által, tanulói korpuszokra kifejlesztett standard alapján. Ez a hallgatók korán és nemén kívül egyrészt a hallgatók nyelvtudásának felmérését, másrészt az ún. nyelvi biográfia felvázolását foglalja magába. Előbbi az e célra kifejlesztett tesztek segítségével sorolja a hallgatói teljesítményeket a Közös Európai Referenciakeretben meghatározott szintekre, ezek közül a B2-es, ill. a C1+ (C1 és afölötti) szintet elérő hallgatók szövegei kerülnek csak annotálásra. A nyelvi biográfia a hallgatók által beszélt nyelveket, ezek sorrendjére, az elsajátítás módjára, ill. a célnyelvi környezetben töltött időre vonatkozó információkat tartalmazza.

4.2 Az annotálási folyamat gyakorlati szempontból és hozzáférhetőség

A gyűjtött anyagok papíralapú és elektronikus archiválását, a kézírásos szövegek begépelését és az annotációt a hallgatói segéderők végzik a projekt mentoráló tagjainak felügyeletével és segítségével. Ez – az annotációt megelőző bevezetésen túl – a szövegek, ill. annotátumok ellenőrzését, a felmerülő kérdések megválaszolását és állandó visszacsatolást jelent. A célhipotéziseket anyanyelvi projekttagjaink lektorálják.

A hibatagek alkalmazásával kapcsolatos problémás eseteket projektmegbeszélések keretében oldjuk meg. Az elmúlt két évben így számos tag revideálásra, ill. megszüntetésre került, például a sok esetben nehezen elhatárolható „kongruencia a determinatívfrázisban" hibakategória, amely a németben gyakori szinkretizmusjelenségek miatt gyakran egybeesett más (melléknév)ragozási hibákkal. Más hibakategóriákat a ritka előfordulás miatt összevontunk, például a különböző (rémakiemelő ill. tagadó) partikulák szórendi hibáit egy egységes tag használatával tesszük kereshetővé. Új hibakategóriaként jelent meg például legutóbb a kötött bővítményeken kívül megjelenő szemantikai viszonyok hibás grammatikai kifejezését kódoló „SemRel" tag.

Az annotátumok feldolgozásának utolsó lépése az ANNIS-kompatibilis verzióba való konvertálás, mely által a szövegek az ANNIS által nyújtott keresési lehetőségek számára hozzáférhetővé válnak (l. 5. pont).

Az annotáció következő szakaszában az új fejlesztésű orig-, graph- és layout-sorok, tehát a szövegeket író megnyilatkozók általi változtatások, ill. az eredeti hallgatói szövegek tördelésének integrálása történik meg az első változatban publikált annotátumokba.

Jelenleg mintegy 63 szöveg (kb. huszonkétezer token) annotációja készült el, melyek a Dulko korpuszának 1.0 verzióját (Dulko-v1.0) fogják képezni. A Falko esszékorpusz (FalkoEssayL2v2.4) méretéhez képest (248 szöveg, 144.619 token) a Dulko mérete talán kicsinek tűnhet, viszont összehasonlítva a negyvennégy különböző anyanyelvvel rendelkezőktől gyűjtött szövegekből álló részkorpuszokkal a Dulko mérete tekintélyesnek mondható.[15] Az ANNIS-kompatibilis verziók egyelőre csak a helyi gépeken installált ANNIS programon keresztül hozzáférhetők, internetes publikálás az év végéig várható.

5 Alkalmazási lehetőségek

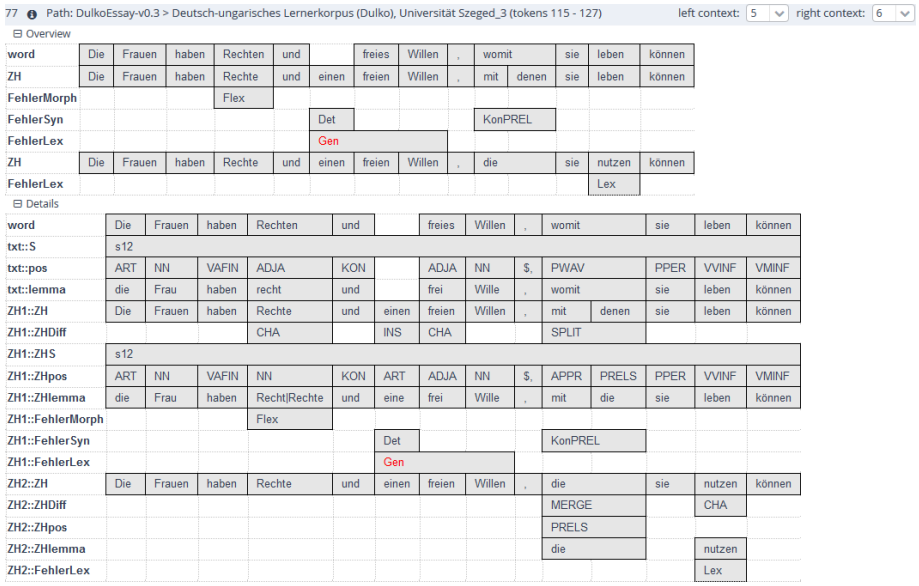
A Dulko annotációs módszerét a Szegedi Tudományegyetem mellett Németországban több egyetemen (Gießen, Lipcse, Marburg és Potsdam), valamint a Genti Egyetemen és Kínában (Hangcsou, Zhejiang University) is használják. Külön örvendetes, hogy a vietnami nyelvtanulói korpusz (az ún. „*Viet*namesisches *L*ernerkorpus", Vielko), amely két vietnami Egyetem (HANU, Hanoi University és a szintén Hanoiban működ-

[15] Egy példával szemléltetve, a Falko esszék szerzői között a negyedik legmagasabb aránnyal képviseltetik magukat a francia anyanyelvűek (vö. Reznicek és mtsai, 2012), az általuk írt szövegek mennyiségénél (17 szöveg, 10.756 token, vö. https://korpling.german.hu-berlin.de/falko-suche/) a Dulko esszéket tartalmazó részkorpusza (34 szöveg, 12.283 token) nagyobbnak bizonyul.

dő ULIS, University of Languages & International Studies) és két németországi egyetem (Lipcse és Gießen) kooperációjában fog létrejönni, ugyancsak a Dulko módszerével készül. A berlini Humboldt-Egyetemen, ahol a Falko korpuszt építették, szintén használják már az EXMARaLDA (Dulko) annotációs szoftvert.

A Dulko a fent említett, közeljövőben építendő korpuszokkal is összevethető lesz, viszont már jelenleg is számos alkalmazási lehetőséget rejt. A következőkben azt kívánjuk szemléltetni, hogy a korpusz milyen kérdésfelvetések tisztázásához járulhat hozzá, egyrészt önmagában, másrészt egyéb korpuszokkal való összehasonlításban.

A fent vázolt annotációs eljárásnak köszönhetően a korábbiakhoz képest egyszerűsödik a különböző hibatípusok elemzése. Ugyan a Falko nyelvtanulói korpuszokban[16] is kereshetők hibatípusok, viszont ez több esetben csak úgy oldható meg, hogy a nagy mennyiségű találati listákból manuálisan választjuk szét a valóban a keresett jelenséghez tartozó, helyesen felismert találatokat (true positives) a hibás, hamis pozitív találatoktól (false positives). A Dulkóban a hibatagek használatával egyszerűen lehívhatók a találatok az egyes hibatípusokhoz, például a Falkóban nehezen kereshető nyelvtani nem tévesztése hibatípusnál a „Gen” hiba-tag segítségével (vö. 8. ábra).



8. ábra Példa az ANNIS keresőfelületéről a „Gen” hibataggel való lekérdezésre kapott találatra

A lekérdezést a metaadatok segítségével is korlátozhatjuk. Így könnyebbé válik különböző tényezők, például a nyelvtanulók nyelvi szintje és az egyes hibatípusok aránya közötti összefüggések vizsgálata. A fenti hibakategória részkorpuszok szerinti megoszlását szemléltetve elmondható, hogy a szintfelmérő nyelvi teszteken a Közös Európai Referenciakeret alapján B2-es nyelvi szinten álló hallgatók a C1-es nyelvi szinten állókhoz képest sokkal több esetben tévesztik el a nyelvtani nemet. Az esszé

[16] https://korpling.german.hu-berlin.de/falko-suche/

részkorpusz aktuális változata (DulkoEssay-v0.3, össz. 12.283 token, 679 mondat) alapján a B2-es részkorpuszban 9871 tokenre (549 mondatra) jut 99 hiba, a C1-es részkorpusz esetén 18 hiba található 2412 token (130 mondat) mennyiségű szövegben.[17] A B2-es részkorpuszban tehát minden 100. tokenre jut egy ilyen hiba, míg a C1-es részkorpuszban csak minden 134.-re. Ugyanígy a célnyelv tanulásával töltött eddigi időtartam és egyéb más metaadatok is könnyen kombinálhatók a hibatagekkel. Így egyszerűbben vizsgálható az a kérdés is, hogy az egyes nyelvi szinteken található hibák mennyiben korrelálnak egymással és mennyiben függnek egyéb tényezőktől, pl. másik tanult idegen nyelv (angol) hatásától.

A magyar anyanyelvű nyelvtanulók esszéit más anyanyelvű nyelvtanulók esszéivel összevetve arról is képet kaphatunk, milyen szerkezetek fordulnak gyakrabban elő az előbbiekben.[18] A Falko nyelvtanulói korpusz esszé részkorpusza (falkoEssayL2v2.4) például mindössze öt találatot tartalmaz 248 szövegben (144.619 token, 6484 mondat) a *solch* (’az a’, ’olyan’) + főnév után álló vonatkozói mellékmondatokra, míg a jóval kisebb Dulko esszékorpuszban (DulkoEssay-v0.3) 12.283 tokenre jut három ilyen szerkezet. A magyar anyanyelvű nyelvtanulók esszéiben tehát minden 4049. tokenre jut egy ilyen szerkezet, míg a Falko korpuszban ez a szám 28.924, azaz 7-szer magasabb. A Falko korpusz német anyanyelvűek által írt esszéiben (falkoEssayL1v2.3) pedig a megfelelő keresőparancs nulla találathoz vezet.[19] Az amúgy a német anyanyelvűek szövegeiben is (bizonyos esetekben) használatos szerkezet tehát a magyar anyanyelvűek szövegeiben sokkal gyakrabban fordul elő (overuse). Ezzel kapcsolatban az egyik lényeges kérdés az interferencia jelensége, azaz hogy milyen esetekben befolyásolják magyar anyanyelvi szerkezetek a célnyelv használata során választott szerkezeteket. Az ilyen egyértelmű eredmények empirikus bizonyítékkal szolgálnak az interferencia hatására. Úgy véljük, a nyelvi hibák elemzésén túl a magyar nyelvtanulók köztes nyelvéről az ilyen kvantitatív elemzések (overuse és underuse) is sokat elárulnak.

Különösen fontos, hogy a nyelvtanítás során tisztában legyünk azzal, milyen jellegű hibák jellemzők a magyar anyanyelvű nyelvtanulók nyelvhasználatára. Ilyen például az *auch* (’is’) fókuszpartikula (rémakiemelő partikula) hibás használata. A magyar anyanyelvű nyelvtanulók annotált esszészövegeiben ez minden hatodik esetben, azaz 93-ból 15 nyolc esetben, – szintén a magyar nyelv hatására – nem a fókusz előtt, hanem azután áll. Az ilyen szórendi hiba a Falko esszékorpuszában (falkoEssayL2v2.4) több mint háromszor ritkább (853 esetből kevesebb mint 45 szórendi hiba tartozik ide, mivel a lemma="auch" & ZH2Diff="MOVS" & #1_=_#2 keresőparancs hamis pozitív találatokat is eredményez). További empirikus vizsgálatokkal a

[17] A keresőparancs a B2-es részkorpusz esetén FehlerLex="Gen" & meta::learner_level_CEFR_conversion="B2", a C1-es részkorpusz esetén pedig FehlerLex="Gen" & meta::learner_level_CEFR_conversion="C1".

[18] A Falko korpusz mellett a Kobalt-DaF projekt (vö. Zinsmeister és mtsai, 2012) nyelvi adataival való összehasonlítás is lehetséges, melyek a Falko korpuszban alulreprezentált kínai vagy svéd anyanyelvű nyelvtanulók nyelvi produktumaival való összehasonlításhoz kínálnak megfelelő alapot.

[19] A Falko esszékorpuszok esetében a lemmára, szófajokra, írásjelre, valamint ezek sorrendjére vonatkozó keresőparancs: lemma="solch" & pos="NN" & lemma="," & pos="PRELS" & #1.#2 & #2.#3 & #3.#4. A Dulko esszékorpusz esetében ez csak annyiban tér el, hogy az újabb lemmatizáló a „solche” szótári alakkal dolgozik a régebbi „solch” helyett.

magyar anyanyelvű nyelvtanulók nyelvhasználatára jellemző hibák azonosítása azért is fontos, mert bizonyos hibák csak nehezen leküzdhetők, ezek sok esetben fosszilizálódhatnak, a haladó szintű nyelvtanulók nyelvhasználatában is megmaradhatnak. Ha ezeknek az általános és középiskolai nyelvtanítás során nem szentelnek kellő figyelmet, a hallgatóknak az egyetemi nyelvoktatás során kell a helyes célnyelvi használatot elsajátítaniuk. A nyelvtanulói korpuszok elemzése tehát az egyetemi nyelvtanítás optimalizálásához is hozzájárulhat. A még haladó szintű nyelvtanulók által is elkövetett (de reményeink szerint az egyetemi évek végére leküzdött) tipikus hibák ismerete az általános és középiskolai nyelvtanárok számára is fontos, hogy diákjaikat minél eredményesebben tudják hozzásegíteni a célnyelv magas szintű használatához.

6 Összegzés

A cikkben bemutattuk a Dulko német-magyar nyelvtanulói korpuszt, amelyben magyar anyanyelvű, németül tanuló germanisztika szakos hallgatók nyelvi adatait annotáljuk és tesszük elektronikusan kutathatóvá. Az annotáció során a Falko korpusz módszerét követjük (pl. szófajok és lemmák automatikus annotációja, metaadatok, célhipotézisek), viszont számos újdonságot vezettünk be (pl. hibatagek használata, fordításszövegek annotációja, a kézirat önjavításainak jelölése). Az EXMARaLDA programra épülő nyílt forráskódú program megkönnyíti a nyers szövegtől az annotált szövegig tartó folyamat technikai megvalósítását és az annotátumok más korpuszokkal való kompatibilitását. A programunkat már jelenleg is több korpuszprojektben használják, és örömmel várjuk a további együttműködéseket.

A Dulko reményeink szerint nemcsak a szabadon kereshető, magyar anyanyelvű németül tanulók nyelvi adataival járul hozzá a nyelvtanulók köztes nyelvének vizsgálatához, hanem a nyílt forráskódú programmal és a javított annotációs eljárással más nyelvtanulói korpuszok építéséhez is mintát és hathatós segítséget nyújt.

Hivatkozások

Brdar-Szabó, R.: Nutzen und Grenzen der kontrastiven Analyse für Deutsch als Fremd- und Zweitsprache. In: Krumm, H.-J., Fandrych, C., Hufeisen, B., Riemer, C. (szerk.) Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch (Sprach- und Kommunikationswissenschaft 352). pp. 518–531. de Gruyter, Berlin, New York (2010a)

Brdar-Szabó, R.: Kontrastive Analyse Ungarisch-Deutsch. In: Krumm, H.-J., Fandrych, C., Hufeisen, B., Riemer, C. (szerk.) Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch (Sprach- und Kommunikationswissenschaft 352). pp. 732–737. de Gruyter, Berlin, New York (2010b)

Fekete, O.: Forschungsmethodologische Aspekte zur Kasusverwendung bei ungarischen DaF-Lernenden. In: Böttger, L., Masát, A. (szerk.) Jahrbuch der ungarischen Germanistik 2008. pp. 163–183. Gondolat, Budapest, Bonn (2009)

Fekete, O.: Komplexität und Grammatikalität in der Lernersprache : eine Längsschnittstudie zur Entwicklung von Deutschkenntnissen ungarischer Muttersprachler. Waxmann, Münster, New York (2016)

Granger, S., Paquot, M.: Core metadata for learner corpora. Draft 1.0. Kézirat. Louvain-la-Neuve (2017)

Gunkel, L., Murelli, A., Schlotthauer, S., Wiese, B., Zifonun, G.: Grammatik des Deutschen im europäischen Vergleich. Das Nominal. Unter Mitarbeit von C. Günther und U. Hoberg. 2 Bände. (Schriften des IDS 14) de Gruyter, Berlin, Boston (2017)

Hirschmann, H., Nolda, A.: Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus. In: Eichinger, L., Plewnia, A. (szerk.) Neues vom heutigen Deutsch: Empirisch – methodisch – theoretisch (Institut für Deutsche Sprache, Jahrbuch 2018). pp. 339–342. de Gruyter, Berlin (2019)

Institut für Deutsche Sprache: Deutsches Referenzkorpus – DeReKo. Archiv der Korpora geschriebener Gegenwartssprache. Institut für Deutsche Sprache, Mannheim (2004ff.) [http://www1.ids-mannheim.de/kl/projekte/korpora]

Juhász, J.: Probleme der Interferenz. Akadémiai Kiadó, Budapest, München (1970)

Krause, T., Zeldes, A.: ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities 31, 118–139 (2016) [http://dsh.oxfordjournals.org/content/31/1/118]

NAT - Nemzeti alaptanterv - Hatály: 2018.I.1. - Magyar joganyagok - a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról. [https://net.jogtar.hu/getpdf?docid=a1200110.kor&targetdate=&printTitle=]

Nolda, A.: Annotation von Lernerdaten mit EXMARaLDA (Dulko). Kézirat. Berlin (2019) [https://andreas.nolda.org/publications/nolda_2019_annotation_lernerdaten.pdf]

Pilarský, J. (szerk.): Deutsch-ungarische kontrastive Grammatik. 2. kiadás. Egyetemi Kiadó, Debrecen (2018)

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., Andreas, T.: Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik – Korpuslinguistik, Berlin (2012) [https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/]

Reznicek, M. Lüdeling, A., Hirschmann, H.: Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In: Díaz-Negrillo, A., Ballier, N., Thompson, P. (szerk.) Automatic treatment and analysis of learner corpus data (Studies in Corpus Linguistics 59). pp. 101–123. John Benjamins, Amsterdam (2013)

Schiller, A., Teufel, S., Stöckert, C., Thielen, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Kézirat. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung und Universität Tübingen, Seminar für Sprachwissenschaft (1999) [http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf]

Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Jones, D. B., Somers, H. L. (szerk.) New Methods in Language Processing, pp. 154–164. Routledge, London (1997)

Schmidt, T.: EXMARaLDA – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache. In: Buchberger, E. (szerk.) Proceedings of Konvens 2004, Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5, Österreichische Gesellschaft für Artificial Intelligence, Wien (2004) [https://www.exmaralda.org/files/Konvens_Paper.pdf]

Selinker, L.: Interlanguage. International Review of Applied Linguistics. Language Teaching 10, 209–231 (1972)

Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). pp. 385–389. European Language Resources Association, Las Palmas de Gran Canaria (2002)

Walter, M., Grommes, P.: Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Niemeyer, Tübingen (2008)

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., Skiba, D.: Das Wissenschaftliche Netzwerk „Kobalt-DaF“. Korpusbasierte Analyse von Lernertexten für Deutsch als Fremdsprache. Zeitschrift für Germanistische Linguistik 40(3), 457–458 (2012)

Nesze semmi, fogd meg jól!

Zéró kopulák automatikus felismerése neurális gépi fordítással

Dömötör Andrea[1,3], Yang Zijian Győző[2,3], Novák Attila[2,3]

[1]Pázmány Péter Katolikus Egyetem Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
[2]Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
[3]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
{domotor.andrea, yang.zijian.gyozo, novak.attila}@itk.ppke.hu

Kivonat Kutatásunkban a nominális mondatok zérókopula-jelenségével foglalkozunk, miszerint bizonyos default esetekben a predikatív névszók önmagukban, testes segédige jelenléte nélkül is betölthetik az állítmányi funkciót. Ennek gépi kezelésére létrehoztunk egy eszközt, amely a zéró kopulás mondatok automatikus felismerésére alkalmas, mi több képes a zéró kopulát a mondatok megfelelő helyére beilleszteni. Az általunk létrehozott eszköz in-domain, azaz a tanítóanyaggal megegyező forrásból származó tesztanyagban közel 90%-os pontossággal képes a zéró kopulák helyes beillesztésére.
Kulcsszavak: zéró kopula, szintaxis, gépi tanulás, gépi fordítás, korpusznyelvészet

1. Bevezetés

A zéró kopula jelensége, miszerint bizonyos default esetekben a predikatív névszók önmagukban, testes segédige jelenléte nélkül is betölthetik az állítmányi funkciót, számos nyelvben ismert. A magyarban a kijelentő mód, jelen idő, 3. személy ilyen default eset.

(1) a. Fábián elvtárs is vámpír **Ø**.
b. Lehetetlen, hogy Fábián elvtárs is vámpír **legyen**.
c. Fábián elvtárs is vámpír **volt**.
d. Én is vámpír **vagyok**.

Kutatásunk célja egy olyan eszköz kidolgozása, amely képes az (1a) típusú mondatok megfelelő helyére beilleszteni a zéró kopulát. „Megfelelő hely" alatt azt a pozíciót értjük, ahol nem default esetben a testes kopula lenne (vö. 1a és c).

Fontos megjegyezni, hogy jelen kutatásban csak azokat tekintjük kopulás mondatnak, ahol az adott default esetben soha nem fordul elő testes ige. Nem soroljuk ide a (2)-típusú létige–semmi váltakozásokat (az alábbi példák forrása az MNSz2 (Oravecz és mtsai, 2015)). Más szóval jelen tanulmány csak a nominális mondatok zérókopula-jelenségével foglalkozik, és nem terjed ki az opcionálisan elhagyható egzisztenciális létigékre (2a-b) vagy a címek szerkezeti sajátosságaira (2c).

(2) a. Ott (van) a csodarendszer, hát alkalmazzák!

b. Ennek semmi értelme (nincs), csak eszembe jutott.

c. Veszélyben (van) az olajellátás.

Célunk elsősorban a korpusznyelvészeti kutatások támogatása: egy olyan eszközt szeretnénk kifejleszteni, amely segítségével a nominális mondatok kvantitatív vizsgálatára alkalmas, nagy méretű korpusz hozható létre. Emellett a zéró kopulás mondatok automatikus felismerése a számítógépes mondatelemzés és szövegfeldolgozás számára is hasznos információ lehet.

2. Kapcsolódó munkák

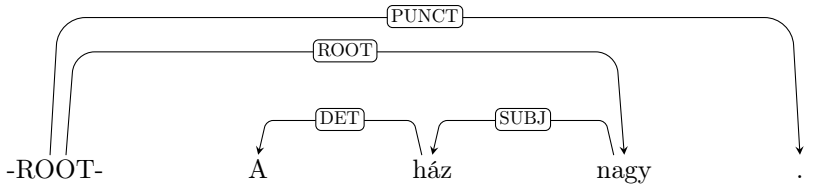
Simkó és Vincze (2017) a zéró kopulás mondatok függőségi elemzésében háromféle megközelítéssel kísérletezik: a funkciófejes, a tartalmas fejes és a komplex címkés elemzéssel. A funkciófej-elemzést követi a Szeged Treebank (Vincze és mtsai, 2010), ahol a névszói állítmány egy üres fejhez kapcsolódik, így a zéró kopulás mondatok a testes kopulásokkal analóg elemzést kapnak. A tartalmasfej-elemzés ezzel szemben nem enged meg üres fejeket, ezért a kopulás mondatok feje minden esetben a névszói állítmány. Ez egyébként megfelel a Universal Dependencies elveinek (Nivre, 2014). A komplex címkés elemzésben a kopula hiányát a névszói állítmány speciális ROOT-VAN-PRED címkéi jelzik. A tanulmány kísérletei során a Bohnet parsert (Bohnet, 2010) tanították be a szerzők a három elemzési módszerre. Eredményeik szerint a parser a funkciófejes elemzést tanulta meg a legsikeresebben, ami azt mutatja, hogy a zéró kopulák beillesztése valóban hasznos a függőségi elemzés számára.

Az üres funkciófejek automatikus beillesztésével kapcsolatos kísérleteket mutat be Seeker és mtsai (2012). Kutatásuk célja a zéró szóalakok (ellipszis miatt hiányzó szavak vagy zéró kopulák) megjóslása a függőségi elemzés során. A cikk három módszert mutat be: az elsőben az üres fejeket a parser illeszti be az elemzés során, a másodiknál az üres fejek a címkekészletben vannak kódolva, míg a harmadik esetén az üres fejek szükségességéről egy osztályozó dönt az elemzés előtt. Számunkra ez utóbbi módszer a leginkább érdekes, hiszen itt a miénkhez hasonló feladatról van szó. A fő különbség, hogy Seeker és mtsai (2012) az üres fejeket a tagmondat elejére illeszti be, és nem oda, ahol a felszíni szerkezetben a zéró kopula helye lenne, így valóban csak osztályozást végez. A másik különbség, hogy a cikkben ismertetett osztályozó elemzett szövegekkel dolgozik, azaz morfológiai információt is használ, míg esetünkben csak az elemzetlen mondatok

állnak rendelkezésre. Csak a zéró kopulás mondatokat tekintve Seeker és mtsai (2012) 83,6%-os pontosságot, 69,2%-os fedést és 75,8-os F-mértéket ért el. A saját módszerünkre is készítettünk ezzel összevethető, csak az osztályozás sikerességét mérő kiértékelést.

Tudomásunk szerint a zéró kopula felszíni szerkezetbe való beillesztésére alkalmas eszköz nem áll rendelkezésre a magyar nyelvre. Az elérhető szintaktikai elemzők sem alkalmazhatók erre a feladatra, ezek ugyanis, a Universal Dependencies elveit követve, nem illesztenek zéró kopulát az elemzéseikbe. Ennek megfelelően az `e-magyar` (Váradi és mtsai, 2018) elemző a zéró kopulás mondatok esetén a tartalmasfej-elemzést alkalmazza (3. példa).

(3)



A korpuszokat tekintve említettük, hogy a Szeged Treebankben vannak zéró kopulát pótló üres fejek. Ezek azonban nem „valódi” mondatrészként, csak virtuális csomópontokként szolgálnak, így a mondatbeli pozícióknak nem tulajdonítottak jelentőséget a korpusz készítői. Az üres fejeket jelölő szimbólumok így gyakorlatilag véletlenszerű helyeken jelennek meg a felszíni szerkezetben, ezért ezeket nem tudjuk közvetlenül zéró kopulás tanítóanyagként felhasználni az általunk kitűzött célra. Előnye azonban a Szeged Treebanknek, hogy bináris osztályozási feladatra (zéró kopulás-e a mondat vagy nem) gold standard adatként rendelkezésre áll. A korpusz 16003 darab zéró kopulás mondatot tartalmaz, ami a teljes méretének nagyjából 17%-a. Ezt az empirikus arányt használtuk fel a tanító- és tesztanyagaink összeállításában.

3. Módszer

Kutatásunkban a feladat megoldásához a gépi fordítás módszerét alkalmaztuk, melynek lényege, hogy transzformációt képez tetszőleges forrás- és célnyelvi mondatok között, ahol a rendszer betanításához nem kell más, mint egy kétnyelvű párhuzamos korpusz.

A gépi fordítás módszerével való megközelítés indokolt, hiszen a forrás- és a célnyelvi mondat azonos, kivéve a zéró kopulás mondatpárokat, melyben a célnyelvi mondatban a zéró kopula helyén egy <zerokop> címke áll.

Munkánk során a Marian NMT (Junczys-Dowmunt és mtsai, 2018) nevű keretrendszert használtuk, ami egy c++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően [1] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

[1] https://marian-nmt.github.io/

Az NMT tanításához a jelenleg „state-of-the-art" Transformer (Vaswani és mtsai, 2017) modellt és Sentece Piece (Kudo és Richardson, 2018) tokenizálót használtuk. A rendszer beállításai és paraméterei a következők:

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0.1;
- learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
- optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
- label-smoothing: 0.1; exponential-smoothing

4. A tanítóanyag

A rendszer tanításához olyan tanítóanyagra van szükség, ahol a zéró kopulák jelölve vannak a mondatokban. Mint említettük, eddig egy ilyen korpusz létezik, a Szeged Treebank, ez azonban egyrészt túl kicsi, másrészt a zéró kopulák helyét tekintve rendszertelen. Elkerülhetetlen volt tehát, hogy új, saját zérókopula-korpuszt hozzunk létre.

Az alapötlet az volt, hogy csináljunk zéró kopulás mondatokat a testes kopulás mondatokból. Ehhez először is arra volt szükség, hogy a testes kopulákat elkülönítsük a lexikális VAN létige azonos alakjaitól (azaz a lokatív, az egzisztenciális és a birtokos igétől). Első kísérletként kipróbáltuk, hogyan oldja meg ezt a feladatot az `e-magyar` automatikus szövegelemző rendszere. Az igék osztályozásához elkészíttettük egy 1000 mondatos teszthalmaz függőségi elemzését, és a kapott elemzés szerint osztályoztuk a kérdéses létigéket. (A teszthalmaz egy random minta volt, 598 mondatban szerepelt benne kopula, és 402-ben lexikális ige.) Ha a kérdéses létigéhez tartozott vele `PRED` relációban álló névszó, akkor *kopula* címkét kapott, egyéb esetben pedig *lexikális*at. A kiértékelés során a kopulás találatok pontosságát és fedését mértük. A módszer 87,8%-os pontosságot és 81,0%-os fedést ért el.

Az eredmények szerint a különböző létigetípusok megfelelő elemzésének kiválasztása nem triviális feladat egy automatikus elemzőrendszer számára. A nehézség egyik oka az lehet, hogy a magyar nem konfigurációs nyelv, így a szórend nem segít abban, hogy megállapítsuk az egyes szavak szintaktikai szerepét. A magyar nyelv másik, automatikus feldolgozást nehezítő tulajdonsága a pro-drop, emiatt nem lehet a létigetípusok megkülönböztetését az esetrag nélküli névszók számára alapozni, hiszen ha csak egyetlen ilyen is van, akkor is könnyen lehetséges, hogy ez az egyetlen névszó predikatív, az alany pedig nincs jelölve.

Alternatív módszerként egy angol-magyar párhuzamos korpuszt használtunk kopulás mondatok gyűjtésére, az angol nyelv konfigurációs jellegének köszönhetően ugyanis az angol mondatokon sokkal könnyebb lokális információk alapján, szabályok segítségével szintaktikai döntéseket meghozni. Vagyis a magyar mondatok angol megfelelői segítenek abban, hogy meghatározzuk a létige aktuális mondatbeli funkcióját (típusát).

Az adatgyűjtéshez egy lemmatizált, morfológiailag elemzett és egyértelműsített, szó szinten megfeleltetett angol-magyar párhuzamos korpuszt használtunk (Novák és mtsai, 2019). Ennek alapja az OPUS OpenSubtitles korpusz (Lison és Tiedemann, 2016), amely összesen 644,5 millió tokenből álló megfeleltetett mondatpárokat tartalmaz. Az angol oldal elemzése a morpha lemmatizálóval (Minnen és mtsai, 2001) és a Stanford taggerrel (Toutanova és mtsai, 2003) történt. A magyar oldalon a PurePos (Orosz és Novák, 2013) és a Humor (Novák, 2014) eszközök végezték el a morfológiai elemzést. Az elemzett szövegekben mindkét oldalon minden eredeti tokent két token reprezentál: (1) a szótő és a fő szófajcímke, illetve (2) az egyéb morfológiai címkék. Az előfeldolgozott mondatokon a fast align program (Dyer és mtsai, 2013) segítségével szó szintű megfeleltetések készültek. Ehhez a morfológiai címkék külön tokenként ábrázolása előnyös, mert így lehetőség van arra, hogy bizonyos, lexikális megfelelővel nem rendelkező szavakat a morfológiai címkékkel kössük össze, például az angol prepozíciókat a magyar oldalon az esetragokat jelző címkékhez.

Az előzőekben leírt párhuzamos korpuszból kiválasztottuk azokat a mondatokat, ahol a magyar oldalon a létige vagy a kopula valamilyen múlt idejű, harmadik személyű alakja (továbbiakban VOLT) szerepel. Ezeket a mondatokat egy szabályalapú algoritmussal címkéztük.

A címkéző algoritmus első lépésben megnézi a VOLT-nak megfeleltetett angol tokeneket. Ha ezek között szerepel nem segédigei *have* vagy expletív *there*, a mondat `lexikális` címkét kap. Ha a VOLT *be*-vel van megfeleltetve, akkor a mondat lexikális igés és kopulás egyaránt lehet, ekkor tehát további vizsgálati lépésekre van szükség. Ha az előbbi tokenek egyike sem szerepelt a VOLT-nak megfeleltetett angol tokenek között, akkor a mondat kikerült a címkézendő anyagból, ekkor ugyanis feltehetően az eredetitől nagyon eltérő fordításról lehet szó, így a mondat angol megfelelője nem megbízható kiindulópontja a címkézésnek.

Ha a VOLT az angol oldalon *be*-nek felel meg, akkor a program megkeresi az angol mondatban azt a „kulcsszót", amely alapján a címkézés megtörténik. A kulcsszó feltételezésünk szerint vagy egy névszói állítmány, vagy egy nem nominatívuszi argumentum (illetve ezek része). Az algoritmus tehát ezeknek az elemeknek a kanonikus pozícióját keresi az angol mondatban (ami a magyar mondat esetén a szórend kiszámíthatatlansága miatt lehetetlen lenne).

A kulcsszó kiválasztásához először is meg kell állapítani, hogy kijelentő vagy kérdő szórendű mondatról van-e szó. Ezt a kérdőszó megléte, illetve az ige pozíciója alapján ellenőrzi az algoritmus. Kijelentő mondatok esetén a kulcsszó a *be*-t követő első olyan token, amely nem tagadószó vagy NP-t módosító elem (például: *very, more*). Az „alkalmasság" megállapítása elsősorban a szófajcímkék alapján történik. (Ld. 4. példa) Eldöntendő kérdés vagy a *what*, *who*, *whose*, *which*, *how* és *why* kérdőszavak esetén a program az előzőekhez hasonlóan jár el, a szórendváltás miatt plusz egy tokent átugorva. (Ld. 5. példa) Az egyéb kérdószóval (pl. *where*, *when* stb.) bevezetett mondatok `lexikális` címkét kapnak.

(4) a. *Régen ez egy kvalitás volt.*
It used to be **a** quality.
b. *Nem volt otthon.*
He was not **at** home.

(5) a. *Mi volt ez a zaj?*
What was that **noise**?
b. *Miről volt szó?*
What was it **about**?

A kulcsszó kiválasztása után az algoritmus megnézi a kulcsszóhoz rendelt magyar tokeneket, és ezek szófaj- illetve morfológiai címkéi alapján megállapítja a kérdéses létige típusát. Ha a kulcsszóhoz tartozik egy nem nominatívuszi esetet jelölő morfológiai címke, akkor a mondat a `lexikális` címkét kapja. Ha a kulcsszónak névelő vagy nominatívuszi névszó felel meg a magyar oldalon, akkor a mondat címkéje *kopula* lesz.

Némely esetben az algoritmus speciális lexikális szabályokat is tartalmaz, a szófajcímkék ugyanis félrevezetőek lehetnek, például az időjárást és egyéb „környezeti helyzeteket" leíró szerkezetek esetén. Ezeknél a kulcsszókeresés értelemszerűen rossz eredményt ad (6). Ezek a szerkezetek ezért lexikális kivételként vannak kezelve egy névszólista alapján, amely az MNSZ2 kollokációkeresőjével készült.

(6) a. ***Sötét** volt és köd.*
It was **dark** and foggy.

A „környezeti kopulás" szerkezetek mellett van még néhány olyan speciális eset, ahol a kulcsszókeresés félrevezető lehet, ezeket leginkább „állandó fordítási különbségnek" nevezhetnénk. Ez alatt azokat a szerkezeteket értjük, amelyek angolul kopulásak, magyarra viszont lexikális VAN-nal fordítjuk őket. Ennek leggyakoribb esete a *being right* 'igaza van' mondat, de ez alá a speciális lexikális szabály alá soroltam a *being lucky* 'szerencséje van', *being necessary* 'szükség van' és a *being ready* 'kész van' szerkezeteket is.

Az algoritmus teljesítményét ugyanazon az 1000 mondatos mintán értékeltük ki, amelyeket az `e-magyarra` alapozott teszthez is használtunk. Az osztályozó 90,8%-os pontosságot és 91,1%-os fedést ért el, azaz jobban teljesített a függőségi elemzésre alapozott módszernél. Az elért pontosság azonban még így sem közelíti meg egy gold standard tanítókorpusz minőségét.

A hibák elemzése során kiderült, hogy a hibás címkék nagy része nem az algoritmusból, hanem valamilyen „külső körülményből" ered. Tipikusan ilyenek például a hibás szófaji címkézés vagy szómegfeleltetés. Szintén gyakori külső hibaforrás volt az angol eredetitől jelentősen eltérő magyar fordítás. Bár az algoritmus igyekszik az ebből származó problémákat kiküszöbölni azzal, hogy figyelmen kívül hagyja az olyan mondatokat, ahol a VOLT-hoz sem *be*, sem *have* nem volt hozzárendelve, ez a megszorítás azonban még mindig sok olyan mondatot „átenged", ahol a mondatszerkezeti vagy akár jelentésbeli különbségek megnehezítik a címkézést.

Ezeket a címkézési hibákat nehezen lehetne elkerülni, így a továbbiakban ezzel a 90%-os pontosságú kimenettel dolgoztunk. A program által kopulásnak

címkézett mondatokban a testes kopulákat egy `<zerokop>` jelre cseréltük, ezek adták a tanítóanyag pozitív példáit (318843 mondat). Ehhez hozzáadtunk az OpenSubtitles korpuszból 1 millió random mondatot (természetesen ügyelve arra, hogy ne legyen átfedés a zérókopulás mondatokkal), és ebből tanítottuk be az alapmodellt. Ez elég jó pontosságot (89,6), viszont gyenge fedést (58,2) produkált, ami nem meglepő, hiszen a negatív példának szánt 1 millió mondatban valószínűleg sok „valódi", jelöletlen zéró kopula volt. Ennek kiküszöbölésére az OpenSubtitles korpuszból kiszűrtük azokat a mondatokat, ahol a *be* valamilyen jelen idejű alakja szerepel nem segédigei funkcióban az angol oldalon harmadik vagy második személyű alannyal (az utóbbinak az esetek nagy részében harmadik személyű a magyar fordítása). Ugyan az angol kopulás mondatok fordítása nem szükségszerűen kopulás a magyar oldalon és fordítva, ezzel a módszerrel viszonylag hatékonyan ki tudtuk szűrni a hamis negatív példák nagy részét a tanítóanyagból. További szűrésképpen a maradékon lefuttattuk az alapmodellt. A továbbiakban azokat a mondatokat (illetve ezek egy részhalmazát) használtuk negatív példának, amelyekbe a modell nem illesztett be zéró kopulát. Végül, az alapmodellt lefuttattuk a kiszűrt (azaz az angol oldalon *be*-t tartalmazó) mondatokon is, így összesen további 161223 darab zéró kopulás mondatot nyertünk.

Az így rendelkezésre álló adatokból a következő modelleket építettük fel[2]:

- **Eredeti szűrt**: ez a modell a címkéző algoritmus segítségével előállított 318843 zéró kopulás, és 1 millió random kiválasztott nem zéró kopulás mondatot tartalmaz. A tanítóanyagba nem kerültek egyszavas, illetve speciális karaktereket tartalmazó mondatok. Ennek a szűrésnek a zajcsökkentés volt a célja.
- **Bővített szűrt**: A pozitív példákhoz hozzáadtuk az alapmodellel előállított zéró kopulás mondatokat (így összesen 477082 zérót tartalmazó mondatunk lett), a negatív példákat pedig újabb 1 millió random nem zéró kopulás mondattal bővítettük.
- **Eredeti javított**: Az eredeti szűrt modell tanítóanyagába visszakerültek az egyszavas mondatok és néhány olyan speciális karakter, amely a Szeged Treebankben gyakorinak bizonyult (pl. §). Ezen kívül javítottunk néhány, a tesztadatok átnézése során feltűnő könnyen kiszűrhető hibát, amelynek altípusait a (7) alatti példákkal szemléltetjük. Az eredeti modell által zéró kopulásnak jelölt mondatok közül kiszűrtük azokat, ahol a magyar oldalon az eredeti mondatban szereplő *volt(ak)* szóalak valamilyen minden esetben hangsúlytalan (de nem enklitikus) összetevő: kötőszó (7a), vonatkozó névmás (7b) vagy névelő (7d, 7e) után következik. Az előbbieket csak a mindig hangsúlyos lexikális (egzisztenciális (7a) vagy birtokos (7c)) *volt* vagy *voltak* igealak követheti, az utóbbiak pedig gyakorlatilag biztosan az 'ex' jelentésű szintén hangsúlyos *volt* melléknév hibás annotációjával keletkeztek. Tulajdonképpen az első eset speciális változata ezen kívül a vesszőt követő vagy mondat eleji *volt* (7f, 7g). Ebben az esetben ugyanúgy általában kizárt a zéró kopulával való helyettesítés (kivéve a beágyazott vonatkozó mellékmondatok esetét). Ahogy a (7) alatti példák is mutatják, az eredeti algoritmus

[2] A modellek elérhetők: http://nlpg.itk.ppke.hu/projects/zerokopula

nagyrészt akkor hibázott így, ha az eredeti mondat múlt idejű névszói állítmányt és hangsúlyos (lexikális vagy melléknévi) *volt*-ot is tartalmazott, és a szóösszerendelési modell az angol kopulát hibásan hozzákapcsolta az utóbbihoz (is). Emellett az is egy érdekes eset, amikor ugyan kopuláról van szó, de ugyanakkor ellipszis is van a mondatban (7h), ezért múlt időben a *volt* a fókuszos kontrasztív szerkezet miatt kötelezően hangsúlyos, azonban jelen időben muszáj lenne egy másik elemnek megjelennie, hogy legyen, amit fókuszálni lehet. Ez a modell így 314607 zéró kopulás, és 1515204 nem zéró kopulás mondatot tartalmaz. (Azaz, ez a tanítóanyag már megfelel a Szeged Treebankből megállapított empirikus aránynak.)

– **Bővített javított**: Az előző anyaghoz hozzáadtuk az alapmodellel generált zéró kopulás mondatokat, és a negatív példákat is kibővítettük a 17%-os aránynak megfelelően, így ez a modell 475830 zéró kopulás, és 2574207 nem zéró kopulás mondatot tartalmaz.

(7) a. De igen, a legtöbb az volt, de **volt** fehér is, valamint ilyen bébifosbarna is.

b. Az a belépőkártya... volt minden, amim **volt**.

c. Házas volt, és **volt** egy fia.

d. Megértjük, hogy a **volt** férje és Ő üzleti társak voltak.

e. A különleges osztag egy **volt** kihallgatótisztje csinálta.

f. Egy szinttel lejjebb voltak a testőrök, **volt** ejtőernyősök vagy idegenlégiósok.

g. Ahhoz képest, hogy fiú vagy, **voltak** jó válaszok.

h. Nem olyan, mint **volt**.

5. Eredmények

A kiértékeléshez kétféle teszthalmazt használtunk. Az egyik ugyanabból a korpuszból készült, mint a tanítóanyag (OPUS), a másik pedig a Szeged Treebankből. Ez utóbbi azt a célt szolgálja, hogy kipróbáljuk, hogy működik a rendszer más domaineken.

Mindkét tesztkorpusz 2000 mondatot tartalmaz, ebből, a 17%-os empirikus aránynak megfelelően, 340 zéró kopulás. A tesztmondatokat ellenőriztük és kézzel javítottuk, ahol szükséges volt, illetve a Szeged Treebank jelölt zéró kopuláit szintén kézzel kellett a helyükre illeszteni. A végeredmény tehát két, méretben és arányban megegyező, de különböző forrásból származó, és ezáltal különböző szövegtípusokat tartalmazó gold standard tesztkorpusz lett.

A Szeged Treebank mondatain tesztelve azonban váratlan nehézségekbe ütköztünk. Egyrészt kiderült, hogy ezek a mondatok olyan karaktereket is tartalmaznak, amelyek nem szerepelnek az első két modell tanítóanyagában (pl. §),

és ezeket a fordító nem tudta kezelni. Emiatt a javított modellekbe visszakerültek a zajcsökkentés céljával kizárt, speciális karaktereket tartalmazó mondatok. Másrészt gondot okoztak a hosszú mondatok is, az OPUS korpuszból készült tanítóanyag mondatai ugyanis – műfaji sajátosságuknál fogva – jellemzően rövidek voltak. Erre a problémára valószínűleg a tanítóanyag hosszú mondatokkal való kiegészítése lehetne megoldás, de a Szeged Treebankből készült tesztanyagon sajnos nem értünk el olyan eredményt, ami lehetővé tette volna, hogy a program outputját tanítóanyagként használjuk.

Mindegyik modellt lefuttatuk mindkét teszthalmazon, és a kimeneten kétféle kiértékelést végeztünk. Az egyik esetén osztályozási feladatnak tekintettük a zéró kopulák beillesztését, azaz csak azt mértük, hogy hány esetben találja el a program, hogy kell-e zéró kopula a mondatba, azt nem értékeltük, hogy jó helyre illeszti-e be azt. A filmfeliratkorpusz viszonylag egyszerű mondatai esetében egyébként szinte minden esetben, ahol nem a referenciával azonos helyre szúrta be a kopulát az algoritmus, helyes az általa javasolt megoldás is. (4. táblázat) A másik kiértékelésnél már csak azokat a zéró kopulákat tekintettük jó találatnak, amelyek a mondat referenciával azonos pozíciójába kerültek. Az eredményeket az 1. és 2. táblázatok tartalmazzák.

	Osztályozás			**Beszúrás**		
	P	**R**	**F1**	**P**	**R**	**F1**
eredeti szűrt	95,5	84,2	89,5	89,7	79,1	84,1
bővített szűrt	94,9	78,3	85,8	88,7	73,2	80,2
eredeti javított	93,6	82,8	87,9	85,6	75,7	80,4
szűrt javított	94,1	76,0	84,1	86,4	69,8	77,2

1. táblázat. Az OPUS tesztkorpusz eredményei a különböző modellekkel

Az in-domain tesztkorpusz esetén az osztályozási feladaton mindegyik modell magas pontosságot ért el. A fedés tekintetében érdekes módon a korpusz bővítése jelentős visszaesést eredményezett. A beszúrási feladatnál mind pontosságban, mind fedésben az eredeti szűrt modell volt a legjobb, 90, illetve 80%-ot közelítő eredménnyel.

	Osztályozás			**Beszúrás**		
	P	**R**	**F1**	**P**	**R**	**F1**
eredeti szűrt	52,9	29,5	37,8	37,8	21,0	27,0
bővített szűrt	59,6	28,5	38,5	45,1	21,5	29,2
eredeti javított	59,4	25,7	35,9	42,3	18,3	25,6
bővített javított	70,5	29,0	41,1	52,4	21,5	30,5

2. táblázat. A Szeged tesztkorpusz eredményei a különböző modellekkel

Más domaineken azonban már egyik modell sem volt ilyen sikeres. Bár az osztályozási feladat pontosságán sokat javított a korpusz bővítése és javítása (részben talán a speciális karakterek visszakerülésének köszönhetően), az eredmény még így is csak 70% körüli. A beszúrást is a bővített javított modell oldotta meg legjobban, 52%-os eredménnyel. A fedés láthatóan mindkét feladatnál és mindegyik modellnél gyenge, a 30%-ot sem éri el. Ennek egyik oka lehet a fentebb említett technikai problémák miatti „veszteség". Másrészt a Szeged tesztkorpusz különböző műfajú szövegei sok olyan mondattípust tartalmaznak, ami a modellek számára ismeretlen, mert a filmfeliratok műfajára egyáltalán nem jellemző.

	Osztályozás			**Beszúrás**		
	P	**R**	**F1**	**P**	**R**	**F1**
eredeti szűrt	77,7	55,0	64,4	68,0	48,2	56,4
bővített szűrt	80,8	51,7	63,1	71,3	45,7	55,7
eredeti javított	81,4	52,4	63,7	70,1	45,1	54,9
bővített javított	85,4	50,9	63,8	73,9	44,1	55,2

3. táblázat. A tesztkorpuszok összesített eredményei a különböző modellekkel

A két tesztkorpusz eredményeit összegezve (3. táblázat) az látszik, hogy a javított modellek a szűrtekhez képest valamivel magasabb pontosságot, de alacsonyabb fedést produkáltak. Az F-mértéket tekintve nincsenek jelentős különbségek a modellek között, a pontosság- és fedésértékek eltérései nagyjából kiegyenlítik egymást.

	OPUS			**Szeged**		
Modell	**helyes**	**valószínűtlen**	**helytelen**	**helyes**	**valószínűtlen**	**helytelen**
Eredeti szűrt	53.9%	7.7%	38.5%	22.2%	22.2%	55.6%
Bővített szűrt	66.7%	6.7%	26.7%	36.4%	9.1%	54.6%
Eredeti javított	63.6%	4.6%	31.8%	50.0%	6.3%	43.8%
Bővített javított	70.0%	10.0%	20.0%	38.9%	11.1%	50.0%

4. táblázat. Helyes-e a zéró kopula modell által javasolt helye, ahol az nem egyezik a referenciával?

6. Hibaelemzés

A hibásan beillesztett zéró kopulák áttekintésekor feltűnt néhány mondattípus, amelyeket a program látszólag következetesen (vagy legalábbis gyakran) elrontott. Ilyenek voltak például az egyszavas mondatok, ezekbe mindegyik modell hajlamos volt hibás zéró kopulákat beszúrni: **Mióta Ø?; *Elnézést Ø!; *Értem*

Ø. stb. Ez érthető azoknál a modelleknél, ahol a tanítóanyagból kiszűrtük az egyszavas mondatokat, de ugyanezeket az outputokat kaptuk az egyszavas szűrő kikapcsolása után is. Úgy tűnik tehát, hogy a rendszernek mindenképpen nehézséget okoz a nagyon rövid mondatok kezelése.

Hasonlóan általános hibajelenségnek bizonyultak a vonatkozó névmással vagy *mint* kötőszóval összekapcsolt összetett mondatok is (8a és 8b). Ennek valószínűleg a zéró kopulák hasonló helyzetben való gyakorisága lehet az oka. Ugyanígy jellemző hiba, hogy a rendszerek minden esetben zéró kopulát szúrnak be az *ez a(z)...* típusú szerkezetekbe, feltehetően szintén azért, mert ezek gyakran valóban zéró kopulásak (8c). A gyakoriság problémakörébe tartozik még a *tagadószó + NP* szerkezetek hibás zéró kopulával való ellátása is (8d).

(8) a. *Hallottam, hogy megtiltottad a belépést a szentélybe, és a boszorkány Ø, aki magához vette a herceget, eltűnt.

b. *Erzsi híreinek valóságtartalma Ø, mint valami méreg, szívódott fel a szervezetébe.

c. *...mert ez Ø az út a hegy másik felére vezetett el...

d. *A mű nem Ø üzletszerű többszörözése és terjesztése a szabad felhasználás körébe tartozik.

Más forrású lehet, de szintén nagyon gyakori hiba a halmozott jelzős szerkezetek hibás értelmezése is (9). Ezekbe mindegyik modell hajlamos az első jelző után zéró kopulát beilleszteni. Ennek az a valószínű oka, hogy az OPUS korpuszra az ilyen leírások nem jellemzőek, így a rendszernek nem volt esélye ezek helyes kezelését megtanulni.

(9) a. *A bennszülött bivalyszerű Ø, fekete nyakizmai kidagadtak az erőfeszítéstől.

b. *A fordulat a második Ø, 1994-es választás után következett be.

Az előbbihez valamelyest hasonló az a hibatípus, amelyet akár pszicholingvisztikai motiváltságúnak is nevezhetnénk, ezek a mondatok ugyanis a szekvenciális feldolgozás egy pontján (vagy csak egy adott tagmondatot tekintve) valóban zéró kopulásnak tűnhetnek. Tipikusan ilyenek az értelmező szerkezetek (10a), illetve az ellipszisek és a koordinált névszói állítmányok (10b-10d).

(10) a. *Kedden a tokiói tőzsde vezető részvényindexe Ø, a Nikkei 225 mintegy 280 pont, azaz 2,7 százalékos erősödést jelzett.

b. *És mi tudna ezen változtatni,... ha nem egy újabb vihar Ø.

c. *Nem szeretnék valami hatalmas villában vagy panellakásban élni, mert az előbbi túl nagy Ø, az utóbbi túl kicsi lenne a család számára.

d. *A Pénzügyi Szervezetek Állami Felügyelete engedélyezte, hogy Szobonya Csaba Zoltán, az OTP Lakástakarékpénztár Rt. igazgatósági tagja Ø, egyben vezérigazgatója legyen.

Végül meg kell említeni a csak a bővített modellekre jellemző hibajelenségeket. Ezekben előfordult a megszólítások zéró kopulás értelmezése, ami az eredeti modelleknél nem merült fel. Továbbá ezeknél a modelleknél találtuk a „legindokolatlanabb" hibákat is, a rendszer hajlamos volt akár ragozott igék után is zéró kopulát beilleszteni. Ez arra utal, hogy a több iterációs tanítóanyag-gyártásnál fennáll a hibák terjedésének és halmozódásának veszélye.

7. Összegzés

Kutatásunk során létrehoztunk egy eszközt, amely a zéró kopulás mondatok automatikus felismerésére és a zéró kopulák mondatba való beillesztésére alkalmas. In-domain tesztkorpusz esetén az eszköz közel 90%-os pontossággal tudta megfelelő helyre beilleszteni a zéró kopulát. A kutatást kiterjesztettük a Szeged Treebankre, amely jelentős mennyiségű a – főleg egyszerű beszélt nyelvi szövegekből álló – tanítóanyagunktól nagymértékben különböző, és jóval bonyolultabb szerkezeteket tartalmazó jogi, irodalmi, illetve sajtószöveget tartalmaz. Ennek következtében ezen a korpuszon jóval gyengébb teljesítményt mértünk. Cikkünk hibaelemzést is tartalmaz, amelyben áttekintettük rendszerünk jellemző hibatípusait.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 számú projekt keretében az FK 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019)

Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. pp. 89–97 (2010)

Dyer, C., Chahuneau, V., Smith, N.A.: A Simple, Fast, and Effective Reparameterization of IBM Model 2. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 644–648. Association for Computational Linguistics (2013), http://aclweb.org/anthology/N13-1073

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)

Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)

Minnen, G., Carroll, J.A., Pearce, D.: Applied morphological processing of English. Natural Language Engineering 7(3), 207–223 (2001), https://doi.org/10.1017/S1351324901002728

Nivre, J.: Nonverbal Predication and Copulas in UD v2. http://universaldependencies.org/v2/copula.html (2014), accessed: 2020-01-04

Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1068–1073. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), aCL Anthology Identifier: L14-1207

Novák, A., Laki, L.J., Novák, B.: Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 63–71. Szeged University, Szeged (2019)

Oravecz, Cs., Sass, B., Váradi, T.: Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében. In: XI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 109–121. Szegedi Tudományegyetem, Szeged (2015)

Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. Incoma Ltd. Shoumen, Bulgaria, Hissar, Bulgaria (2013)

Seeker, W., Farkas, R., Bohnet, B., Schmid, H., Kuhn, J.: Data-driven dependency parsing with empty heads. In: Proceedings of COLING 2012: Posters. pp. 1081–1090. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012), https://www.aclweb.org/anthology/C12-2105

Simkó, K.I., Vincze, V.: Hungarian copula constructions in dependency syntax and parsing. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). pp. 240–247. Linköping University Electronic Press, Pisa,Italy (Sep 2017), https://www.aclweb.org/anthology/W17-6527

Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 173–180. NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1073445.1073478

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010. ELRA, Valletta, Malta (May 2010)

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (szerk.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)

A duplakocka modell és az igei szerkezeteket kinyerő „ugrik és marad” módszer nyelvfüggetlensége, valamint néhány megjegyzés az UD annotáció univerzalitásáról

Sass Bálint

MTA Nyelvtudományi Intézet, ELTE BTK
sass.balint@nytud.hu

Kivonat Jelen tanulmány egy módszernek a magyartól különböző nyelvekre való alkalmazhatóságát vizsgálja. A (Sass, 2019) tanulmány egy valódi igei szerkezetek kinyerésére szolgáló eljárást mutat be magyar nyelvre, és két állítást fogalmaz meg mellékesen: (1) a módszer tetszőleges nyelvre alkalmazható; (2) a módszer alkalmazásához szükséges adatok függőségileg elemzett korpuszból könnyen származtathatók. E két állítást vesszük górcső alá. Adatként universal dependencies (UD) korpuszokat használunk fel. Az UD-nek köszönhetően annotációs különbségek elvileg nincsenek nincsenek a különféle nyelvű korpuszok között, csak a nettó nyelvi különbségek láthatók. Ezzel kapcsolatban gyakorlati megfigyeléseink alapján kritikát fogalmazunk meg. Bár az ige és közvetlen bővítményei közötti viszonyokat különböző nyelvek különböző eszközökkel fejezik ki, a vizsgált nyelvekre ezek a nyelvi eszközök néhány általános módon megragadhatók: esetrag, elöljáró/névutó (esetraggal vagy anélkül), szórend. Az említett eljárás működésének egyetlen feltétele az igék és közvetlen bővítményeik közötti viszonyok leírása, a fentiek alapján tehát működtethető az algoritmus. Eredményként valódi igei szerkezeket kapunk, azaz az eredmények igazolják sejtésünket, az eredeti cikk állításai megállják a helyüket.
Kulcsszavak: igei szerkezet, valódi igei szerkezet, duplakocka, korpuszháló, ugrik és marad, többnyelvű, universal dependency

1. Motiváció – kiáltvány a szerkezetekért

A nyelv alapegységei nem a szavak, hanem a szerkezetek. A szó csak a szerkezet szélső esete: olyan szerkezet, ami egy elemből áll.

A legegyszerűbb egyszavas kifejezés esetében is nagyon gyakran előfordul, hogy egy másik nyelven a megfelelője többszavas. Azt gondolnánk, hogy ‘*krumpli*’ minden nyelven egy szó, franciául mégis ‘*pomme de terre*’ (földi alma). A szenegáli wolof anyanyelvű beszélők hihetik, hogy az olyan köznapi dolgokra, mint a ‘*gëmm*’ nyilván minden nyelv külön szót használ, magyarul mégis így mondjuk ezt: ‘*behunyja a szemét*’. Azt mondhatjuk, hogy szerencse, ha valamire épp van egyszavas kifejezés egy nyelvben, tetszőleges nyelven lehetséges, hogy a szóban forgó dologra csak többszavas egység, szerkezet létezik.

Másfelől, olyan is gyakran előfordul, hogy egy szó megfelelője egy másik nyelven kötött morféma, ahogy ezt az angol '*in*' és a magyar '*-bAn*' rag egyszerű példája mutatja. Ennek megfelelően a '*believe in*' többszavas szerkezet, míg a '*hisz -bAn*' -ről ez nem mondható el a szó szoros értelmében. Utóbbi inkább csak másfél szavas.

Sőt, bizonyos szerkezeti elemek egyáltalán nem is jelennek meg a felszínen, miközben nagyon is fontosak az adott kifejezés szempontjából. Az angol ditranzitív szerkezetek három eleméről – az alanyról, a direkt tárgyról és az indirekt tárgyról – kizárólag a szórendből tudjuk meg a szerkezetben betöltött szerepüket. Innen nézve a '*give*' egy összetett, négy elemből álló szerkezet: '*give SUBJ OBJ IOBJ*'.

Annak idején az első szótárírók mégis a szavakat kezdték el listázni, első ránézésre a szavak tűntek természetes alapegységnek. Ez a hagyomány azóta is él. A szótárakban címszavakat találunk akkor is, ha egyre inkább teret kap a címszavakhoz kapcsolódó különféle típusú szerkezetek, frázisok bemutatása (Atkins és Rundell, 2008).

Amellett érvelünk tehát, hogy az lenne az üdvös, ha nem szótárakat, hanem *szerkezettár*akat hoznánk létre. A szerkezetek legtöbbször egyértelműsítik a bennük szereplő szavakat, de legalábbis csökkentik a többértelműségüket (Yarowsky, 1993; Pustejovsky, 1995). Ahhoz képest, hogy egy forrásnyelvi igéhez felsorolunk 8-10 igét a célnyelven (Kilgarriff, 1997), sokkal hasznosabb, ha az ige szerkezeteit vesszük számba, és a megfelelő szerkezeteket adjuk meg a másik oldalon.

Az angol '*go*' esetében első körben (például egy kezdő nyelvtanuló számára) elegendő az alábbi három szerkezet ismerete:

- '*go to NOUN*' = megy valahová
- '*going to VERB*' = fog csinálni vmit
- '*go ADJ*' = válik vmilyenné

Az, hogy a szerkezeteket tekintjük alapelemnek az első lépés a címszavak helyett „címszerkezeteket" tartalmazó szerkezettárak megalkotása felé.

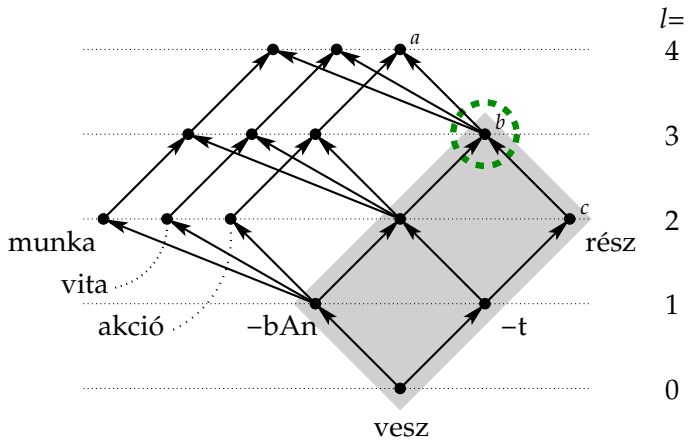
2. Korábbi munka

Kiindulópontunk (Sass, 2019), mely egy függőségileg elemzett korpuszból valódi igei szerkezeteket kinyerő algoritmust mutat be magyar nyelvre. A módszer alapját képező ún. „duplakocka" modellt (Sass, 2018) írja le részletesen.

A valódi igei szerkezet fogalma itt a lehető legáltalánosabb módon magában foglalja a lexikográfiailag hasznos valamennyi igei szerkezetet: a vonzatos igéket ('*vki hisz vmiben*'), az igei szókapcsolatokat ('*süt a nap*') és a vonzatos komplex igéket ('*vki részt vesz vmiben*') is. Éppen ezek azok az egységek, melyeknek egy ige vonatkozásában egy szerkezettárban szerepelniük kell, ezért hasznos ez az összegyűjtésükre szolgáló automatikus módszer.

A modell két kulcseleme az egy tagmondatot és a benne rejlő igei szerkezeteket (azaz az igét és a mellette lévő helyeket és kitöltőket, azaz slot-okat és filler-eket) reprezentáló duplakocka, valamint a korpusz ugyanazon igét tartalmazó összes tagmondatát egyben reprezentáló korpuszháló, mely a duplakockák

egyfajta egymásra vetítésével, kombinálásával áll elő. A korpuszháló megjeleníti az adott ige mellett előforduló összes hely és kitöltő gyakorisági viszonyait (1. ábra).



1. ábra: Az ábra a duplakocka és a korpuszháló fogalmát illusztrálja, valamint bemutatja az „ugrik és marad" módszer működését. Az ábrán látható kicsi korpuszháló a '*részt vesz munkában*', '*részt vesz vitában*' és '*részt vesz akcióban*' tagmondatok duplakockájának kombinációjaként áll elő. A gyakoriságot képviselő f függvény értéke a szürke hátterű csomópontok esetében 3, a többi csomópont esetében 1. Az l az adott csomópontokhoz tartozó szerkezetek hossza. A könnyebb áttekinthetőség kedvéért az alanyi dimenziót nem ábrázoljuk.

A modellen dolgozó algoritmus feladata kettős: meg kell állapítania, hogy mely helyek részei a szerkezetnek és ezek közül hol inherens elem a kitöltő is. A '*vki részt vesz vmiben*' valódi igei szerkezethez 3 hely (alany, tárgy és -bAn) valamint a tárgyi helyet fixen kitöltő '*rész*' elem tartozik elengedhetetlenül hozzá. Tekintsük az adott csomópont által képviselt igei szerkezet korpuszgyakoriságát megadó f függvényt a korpuszhálón. A kidolgozott „jump and stay" („ugrik és marad") módszer arra a megfigyelésre épül, hogy a valódi igei szerkezeteket képviselő csomópontokra igaz az, hogy az f függvény értéke a csomópont fölötti élen jelentősen megnő, a csomópont alatti élen pedig nem változik. A 1. ábrán látható példán azt látjuk hogy az a-b élen a függvény értéke 1-ről 3-ra nő, a c-b élen pedig nem változik. Ez jelzi, hogy a b pontban egy valódi igei szerkezetet találunk. Megfigyelhető, hogy az „ugrás" (a-b) során egy esetleges elemet ('*akció*') hagyunk el, a „maradás" (c-b) során pedig egy szükséges elemet ('*-bAn*') veszünk hozzá. Így a köztes b pont éppen a megkívánt elemeket fogja tartalmazni, így valódi igei szerkezetet kapunk. Vegyük észre, hogy az ábra tetszőleges csomópontjából indulva minden esetben a bekarikázott – helyes – csomóponthoz vezetnek az ugrás+maradás lépéssorozatok.

3. UD korpuszok előfeldolgozása

A módszert eddig kizárólag magyar nyelvű szövegen használták. Az említett cikk két állítást fogalmaz meg: (1) a módszer tetszőleges nyelvre alkalmazható; (2) a módszer alkalmazásához szükséges adatok függőségileg elemzett korpuszból könnyen származtathatók. Jelen tanulmány e két állítást vizsgálja: igyekszik megmutatni, hogy a modell és az algoritmus is nyelvfüggetlen, valamint felméri, hogy mennyi munkával lehet előállítani függőségileg elemzett korpuszból a kívánt bemenetet. Ha reményeink beigazolódnak, az megnyitja az utat tetszőleges, akár kisebb, kevesebb erőforrással bíró nyelvek alapvető szerkezeteinek számbavétele előtt. (Kis nyelvekre természetesen kisebb az esélye, hogy függőségi elemző vagy függőségileg annotált korpusz rendelkezésre áll. De ez nem is elengedhetetlen feltétel. A függőségileg elemzett korpusz kényelmes lehetőséget biztosít a szükséges bemenet előállítására, de egy egyedi szabályalapú eljárás is megfelelő lehet erre a célra.)

Azért bízhatunk a nyelvfüggetlenségben, mert lényegében pusztán arra van szükség, hogy az adott nyelvben legyenek predikátumok, a predikátumoknak argumentumai, és a kettő között valamiféle megragadható viszony. Arra pedig, hogy az inputot egyszerűen elő tudjuk állítani, a szabadon hozzáférhető, egységes annotációval bíró, kézzel annotált, gold sztenderd UD korpuszok (Nivre és mtsai, 2019) adnak lehetőséget[1].

Az UD korpuszok közül a vizsgálatainkhoz cseh, német, angol, finn, magyar, holland, norvég, török és wolof nyelvű korpuszt választottunk[2]. Az elvégzett munka legnagyobbrészt a korpuszok előfeldolgozásából állt. A megfelelő bemenet előállítása után eredeti formájában futtattuk az „ugrik és marad" módszert a valódi igei szerkezetek kinyerésére.

Az előfeldolgozás feladata tehát az igék, valamint az ige közvetlen bővítményeit képviselő helyek és kitöltők meghatározása volt. Ez nagyon hasonló a „konstituensfa felsőszintű szintaktikai elemei"-hez („top level syntactic sequence of the constituent tree") (Shi és mtsai, 2016), azzal a különbséggel, hogy az elemek sorrendjét mi nem vesszük figyelembe.

Az igét és közvetlen bővítményeit a tagmondat tartalmazza, de tagmondatra bontásra a függőségi elemzésnek köszönhetően nem volt szükség. Sőt, úgy döntöttünk, hogy nemcsak a tagmondatok főigéjét, hanem minden egyes igei alakot (UD: UPOS=`VERB`) gyökérnek tekintünk, így a potenciális igei szerkezetek száma megnőtt, és adott igei alak két szerkezetnek is része lehet, az egyiknek gyökérként, a másiknak bővítményként. A '*He didn't think he needed to know anything about South Asia.*' mondatban a '*need*' és a '*know*' is ilyen kettős szerepű ige.

[1] A használt UD terminusok feloldása a `http://universaldependencies.org` oldalon található meg.

[2] Konkrétan az UD 2.4-es verziójából vett alábbi fájlokkal dolgoztunk: cs_pdt-ud-dev.conllu, de_hdt-ud-dev.conllu, en_ewt-ud-train.conllu, fi_tdt-ud-train.conllu, hu_szeged-ud-train.conllu, nl_alpino-ud-train.conllu, no_bokmaal-ud-train.conllu, tr_imst-ud-train.conllu, wo_wtb-ud-train.conllu. Ezek nagyjából egyforma méretű 2-300000 szavas korpuszok.

Ennek köszönhetően tehát az igei szerkezetekben főnévi igenévi bővítmények is megjelennek.

1. hely (slot) megállapítása ♦ Az ige és közvetlen bővítményei közötti viszonyok a vizsgált nyelvekre néhány általános módon megragadhatók: esetrag, elöljáró/névutó (esetraggal vagy anélkül), szórend. Az ige közvetlen dependenseként megjelenő `nsubj`, `obj`, `iobj`, `obl`, `case` és `xcomp` relációval kötődő elemeket vesszük tekintetbe, valamint azokat, melyek tetszőleges reláció mellett rendelkeznek `Case` feature-rel. (Az `xcomp` a fent említett esetet jelenti, mikor a bővítmény egy tagmondat, így saját igéje van általában főnévi igenévi alakban.) Fontos kiemelni, hogy ezen kívül figyelembe vesszük ezen dependensek dependenseként megjelenő elöljárók/névutók (UD: UPOS=`ADP`) lemmáját is. Példa: az '*Acc=in*'[3] olyan közvetlen bővítményi helyet jelöl, amely tárgyesetben áll és van egy '*in*' elöljárója.

Gondot jelent, hogy a német korpuszban az elöljáró+névelő kontrakciók (pl.: '*am*'='*an*'+'*dem*', '*ins*'='*in*'+'*das*') lemmája sajnos megegyezik a szóalakkal, ahelyett, hogy az eredeti elöljáró lenne a lemma. Itt egyedi eljárással mappelni kellett a kontrakciókat az elöljárókra, hogy a szerkezetekben ne váljon ketté a sima elöljáró és a kontrahált forma. Egy másik probléma a főnévi igenevekhez kapcsolódik. Bizonyos nyelvekben a főnévi igenévhez tartozik egy elöljárókhoz hasonló plusz szócska: angolban például '*to*', hollandban '*te*', wolofban '*ci*'. Ez a nagyon specifikus elem az UD annotációban összemosódik más jellegű elemekkel: szófaja partikula (UPOS=`PART`), hasonlóan a cseh '*je*' (csak), norvég '*ikke*' (nem) szóhoz vagy a magyar '*meg*' igekötőhöz; függőségi relációja `mark`, ami pedig az összes alárendelő tagmondatot jelölő elem közös kódja. Emiatt ezek az elemek végül is csak nyelvfüggő módon, a szóalakjuk alapján ragadhatók meg. A főnévi igenév jelölőszócskája egy olyan egyedi elem, amelynek érdemes lenne bevezetni egy külön egyedi szófajt/kódot, amit nagyon jó lenne az összes korpuszban egységesen használni.

2. kitöltő (filler) megállapítása ♦ A kitöltő az ige közvetlen dependensének kisbetűsített lemmája lesz. Sass (2019) említi, hogy a névmások, mivel nagyon gyakoriak, hajlamosak megjelenni kitöltőként, pedig általában nincs idiomatikus jelentésük. A cikk javaslata szerint a névmásokat (UD: UPOS=`PRON`) az előfeldolgozás során töröljük, kivétel ezalól a '*maga*' és az '*egymás*'. A '*maga*' megfogható a `Reflex=Yes`, az '*egymás*' pedig a `PronType=Rcp` UD feature alapján.

Gondot jelent, hogy a '*maga*' esetén a német '*sich*' annotációja eltér ettől, így külön kell kinyerni a lemmája alapján. Az '*egymás*' esetén összetettebb a helyzet, a cseh '*navzájem*', a német '*einander*' és a török '*birbiri*' esetén különféle eltérő módokon jelölik az annotációban a korpuszok ezt a szót. Az '*einander*' elöljárós alakjai további problémát jelentenek: ezeket az alakokat egybeírjuk ('*miteinander*', '*zueinander*'), a korpuszban úgy döntöttek, hogy ezeket az egybeírt alakokat adják meg lemmaként (!) ahelyett, hogy az eredeti szó lenne a lemma, és

[3] '*Acc=in*': itt az egyenlőségjel két elem összetartozását jelöli, azt, hogy az adott helyet két elem együtt reprezentálja.

külön elemként kapcsolódna hozzá az elöljáró – függetlenül attól, hogy egybeírjuk. Még további problémát jelent az angol '*each other*', ez esetben a külön két szóba írás a gond. Ez a kifejezés teljesen önálló kódot kap (`DET+ADJ`), külön egyedi megoldással lehet csak rátalálni. A főnévi igenév jelölőszócskájához hasonlóan az '*egymás*' is tipikus esete az olyan különleges szónak, aminek saját szófaj, saját kód dukál, amit aztán az összes korpuszban egységesen lehet használni.

3. ige megállapítása ♦ Az ige kisbetűsített lemmája elé kapcsoljuk az esetleges elváló igekötőt. Az igekötő-ige kapcsolatot általában a `compound:prt` UD reláció jelzi. Az egységesség kedvéért minden nyelvben igekötő–ige sorrendben szerepeltetjük az igekötős igék elemeit, ez az angolban '*upbreak*', '*inturn*' alakokat eredményez.

Gondot jelent, hogy a magyar korpuszban erre a relációra eltérő jelölést (`compound:preverb`) használnak. Szintén probléma, hogy az angol korpuszban ugyanarra a jelenségre többféle annotáció használatos: a '*stir up*' például helyesen `compound:prt`, a '*get through*' vagy a '*go away*' viszont `advmod` (`ADV` szófajjal). A vizsgált 9 nyelv közül háromban (cseh, finn, török) nem találtam elváló igekötőt. Lehetséges, hogy van, csak ismét más a jelölés. További eltérés, hogy a magyarban + kapcsolja össze az egybeírt igét és igekötőt, más nyelvekben (pl.: német) nem jelzi ezt semmi. A holland '*plaatsnemen*' igekötős ige, míg formai és egyben tartalmi angol megfelelője '*take place*' ige+tárgy szerkezetű.

A universal dependency treebank-ek kiváló erőforrások, bár éri némi kritika is őket (Osborne és Gerdes, 2019). A fentiek alapján megállapíthatjuk, hogy nem felelnek meg maradéktalanul annak az alapvetőnek vélt követelménynek, hogy ugyanazon jelenséget mindig ugyanúgy jelöljünk, eltérő jelenséget pedig mindig eltérően jelöljünk („use the same term ... for the same function") (Croft és mtsai, 2017). Azaz mindig minden egységesen, ugyanúgy működjön, hogy amennyire csak lehet, ne kelljen nyelvfüggő lépéseket végezni. Ez a hiányosság legfőképpen azért baj, mert veszélyezteti a „minden találatra szükség van" elvet. Eszerint mindenfajta korpuszkereséskor – az igék, helyek és kitöltők fent részletezett megállapítása is ilyen – a felhasználó mindig az összes találatot szeretné látni, azaz a recall az, ami itt kiemelten fontos. Az UD treebank-ek ezzel együtt nagyon jól használhatók, az előfeldolgozás során a fent részletezett problémákat megoldva igyekeztünk a találatvesztés esélyét a lehető legkisebbre szorítani.

Úgy is fogalmazhatunk, hogy valójában nem „formai" hanem „funkcionális" függőségekre van szükségünk az igei szerkezetek megragadásához, és a treebankekre épülő eljárásokban általában is ezek tűnnek igazán hasznosnak. A fenti átalakító lépések mindegyike tekinthető egy ebbe az irányba – a funkcionális függőségek felé – tett lépésnek, ahol az azonos *funckiójú* elemek, szavak, illetve relációk kapnának azonos jelölést.

A korpuszokban a legalább 20× előforduló igéket vizsgáltuk. Az előfeldolgozó szkriptek és az eredményfájlok elérhetők a `https://github.com/sassbalint/double-cube-jump-and-stay-multilingual` címen. Jelen cikk az `5dde1d7` commit azonosítójú verzióval készült.

4. Eredmények

Az eredményül kapott szerkezetek túlnyomó többsége megfelelő valódi igei szerkezet. Az 1. táblázatban egy mutatvány látható különféle szerkezetekből.

#	nyelv	igei szerkezet	magyar megfelelő
1.	cs	'být SUBJ:rozdíl mezi'	(van különbség vmi között)
2.	cs	'investovat do'	(befektet vmibe)
3.	cs	'stát se OBJ'	(válik vmivé)
4.	de	'fallen SUBJ:aktie auf'	(esik részvény vmire)
5.	de	'finden sich SUBJ:information auf'	(megtalálható információ vhol)
6.	de	'handeln sich um'	(arról van szó)
7.	en	'do IOBJ OBJ:favor'	(szívességet tesz vkinek)
8.	en	'get in touch with'	(kapcsolatba lép vkivel)
9.	en	'make sure'	(meggyőződik)
10.	en	'take OBJ:care of'	(vigyáz vmkire)
11.	fi	'ottaa Ill:huomio OBJ'	(figyelembe vesz vmit)
12.	fi	'ottaa Ill:käyttö OBJ'	(használatba vesz)
13.	fi	'ottaa Ill:käsi OBJ'	(kézbe vesz vmit)
14.	hu	'lesz SUBJ:szükség -rA'	
15.	hu	'tesz lehetővé -t'	
16.	nl	'zien OBJ:kans te'	(lát lehetőséget vmit csinálni)
17.	no	'få OBJ med seg'	(magával visz vmit)
18.	no	'få OBJ:gjennomslag i'	(áttörést ér el vmiben)
19.	no	'få OBJ på seg'	(felvesz vmit (ruhaféleséget) magára)
20.	no	'få OBJ:tillit'	(önbizalma lesz)
21.	no	'få OBJ i løp'	(futtat vmit (szoftvert))
22.	no	'ha OBJ på seg'	(vmi (ruhaféleség) van rajta)
23.	wo	'am OBJ:kàttan ci'	(van energiája vmit csinálni)
24.	wo	'wax IOBJ OBJ'	(mond vkinek vmit)

1. táblázat. Mutatvány a kapott szerkezetekből. A kitöltetlen alanyi helyet nem tüntetjük fel.

Jópár igénél az összes kijövő szerkezet jó. '*look*' → '*look*', '*look good/great*', '*look like*'; a '*deal*' egyetlen szerkezete a '*deal with*'; a '*go*' esetén lényegében azok a szerkezetek jöttek ki, amelyeket korábban említettünk (2. oldal).

Az eredményekből kontrasztív tanulságokat is le lehet vonni. A 2. táblázatban a '*beszél -rÓl*' szerkezet látható az egyes nyelveken. A magyar kivételével valamennyi szerkezetet az „ugrik és marad" algoritmus futtatásának eredményeként kaptuk. A cseh '*čekat*' (vár) összes szerkezete helyes: '*čekat*', '*čekat Acc=na*' (vár -rA), '*čekat OBJ*' (vár -t), '*čekat se*' (várandós). A neki megfelelő német '*warten*' igénél hasonló szerkezeteket kapunk: '*warten*', '*warten Acc=auf*'. Érdekes kérdés, hogy mennyire feleltethetők meg egymásnak a különböző nyelvek preopozíciói. A cseh '*mít OBJ Dat=k:dispozice*' és a német '*stehen OBJ Dat=zu:Verfügung*' (rendelkezésre áll) párhuzama arra utal, hogy a '*Dat=k*' és

hu	beszél	-rÓl
cs	mluvit/hovořit	o
de	sprechen	von
en	talk	about
fi	puhua	Ela
nl	praten	over
no	snakke	om

2. táblázat. A '*beszél -rÓl*' szerkezet megfelelői. Látjuk, hogy a két finnugor nyelv eseteket használ, az indoeurópai nyelvek pedig különféle elöljárókat.

a '*Dat=zu*' megfelel egymásnak. Ezt alátámasztja a '*patřit Dat=k*' – '*gehören Dat=zu*' (tartozik vmihez) pár is.

Az UD korpuszok viszonylag kis méretűek, ezáltal sok esetben nem kiegyensúlyozottak. A német korpusz szerkezetei arra utalnak, hogy a szövegei főként számítástechnikai területről származnak: '*arbeiten unter windows*', '*laufen mit mhz*', '*laufen unter mac*', '*laufen auf system*', '*laufen unter windows*'.

Az implementált névmástörlés jól működik, nyilvánvalóan nincsenek személyes stb. névmást tartalmazó szerkezetek, ugyanakkor a visszaható névmásos szerkezetek jelentős számban megjelennek (ld. pl. az 1. táblázat 3., 5., 6., 17., 19. és 22. szerkezetét). Az „ugrik és marad" módszer ismert korlátai megjelennek: előfordulnak a szerkezetekben gyakori, jellegzetes de nem idiomatikus szavak kitöltőként. A norvég '*få øye på*' (pillantást vet vmire, meglát vmit) kifejezés például '*få øye på løve*' (meglátja az oroszlánt) formában jelenik meg elsősorban a kicsi, az eredeti cikkhez képest két nagyságrenddel kisebb korpuszméret miatt.

5. Összefoglalás

Jelen cikkben egy eredetileg csak magyar nyelvre alkalmazott valódi igei szerkezeteket kinyerő algoritmus – az „ugrik és marad" módszer – nyelvfüggetlenségét vizsgáltuk meg. A módszer csupán a predikátum-argumentum struktúra meglétét követeli meg, így remélhető volt, hogy szinte bármely nyelvre működőképes lesz. Nyolc európai nyelv függőségileg elemzett UD korpuszából nyertük ki az algoritmus bemenetéhez szükséges adatokat. Az UD korpuszok előfeldolgozása során jónéhány helyen ütköztünk a korpuszok nem teljesen egységes, nem teljesen univerzális annotációjából adódó problémákba. Ezeket részletesen elemeztük. Az algoritmus lefuttatása révén helyes valódi igei szerkezeteket kaptunk felügyeletlen módon. Elmondhatjuk, hogy az absztraktban felvetett mindkét állítás megállja a helyét: viszonylag egyszerűen elő lehet állítani függőségileg elemzett korpuszból az algoritmus bemenetét; valamint hogy az algoritmus valóban lexikográfiailag is hasznos valódi igei szerkezeteket szolgáltat számos nyelven, függetlenül attól, hogy az eredeti tanulmányhoz képest két nagyságrenddel kisebb korpuszokkal dolgoztunk. A jövőben tervezzük a módszer nagyobb elemzett korpuszokon való kipróbálását. Eredményeink megteremtik az

alapját annak, hogy szinte tetszőleges, akár kisebb, kevesebb erőforrással bíró nyelvek tipikus igei szerkezeteit összegyűjtsük. A kód és az eredmények elérhetők `https://github.com/sassbalint/double-cube-jump-and-stay-multilingual` címen.

6. Köszönetnyilvánítás

A kutatást az MTA Bolyai János Kutatási Ösztöndíja támogatta (ügyszám: BO/00064/17/1; időtartam: 2017-2020). Az Információs és Technológiai Minisztérium ÚNKP-19-4 kódszámú Új Nemzeti Kiválóság Programjának szakmai támogatásával készült.

Hivatkozások

Atkins, B.T.S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press (2008)

Croft, W., Nordquist, D., Looney, K., Regan, M.: Linguistic typology meets universal dependencies. In: Dickinson, M., Hajic, J., Kübler, S., Przepiórkowski, A. (szerk.) Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15). pp. 63–75 (2017)

Kilgarriff, A.: "I dont't believe in word senses". Computers and the Humanities 31(2), 91–113 (1997)

Nivre, J., Abrams, M., Agić, Ž., et al.: Universal Dependencies 2.4, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University (2019), http://hdl.handle.net/11234/1-2988

Osborne, T., Gerdes, K.: The status of function words in dependency grammar: A critique of universal dependencies (UD). Glossa: A Journal of General Linguistics 4(1), 17 (2019)

Pustejovsky, J.: The generative lexicon. Cambridge, MA, US: The MIT Press (1995)

Sass, B.: Az igei szerkezetek algebrai struktúrája, avagy a duplakocka modell. Argumentum 14(1), 12–44 (2018)

Sass, B.: The 'jump and stay' method to discover proper verb centered constructions in corpus lattices. In: Proceedings of RANLP 2019. pp. 1076–1084. Varna, Bulgaria (2019)

Shi, X., Padhi, I., Knight, K.: Does string-based neural MT learn source syntax? In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1526–1534. Association for Computational Linguistics, Austin, Texas (Nov 2016), https://www.aclweb.org/anthology/D16-1159

Yarowsky, D.: One sense per collocation. In: Proceedings of the workshop on Human Language Technology. pp. 266–271. Princeton, New Jersey (1993)

Egy emBERT próbáló feladat

Nemeskey Dávid Márk[1]

[1] Számítástechnikai és Automatizálási Kutatóintézet
nemeskey.david@sztaki.hu

Kivonat Az utóbbi egy-két évben a mély, kontextuális szóbeágyazások kiszorították a hagyományos, kézzel összeállított feature halmazokat a legtöbb nyelvi feladatban. Ennek ellenére a magyar nyelvfeldolgozó rendszerek (`e-magyar`, `magyarlanc`) még mindig a hagyományos, kézi feature-ökkel dolgoznak. A cikkben bemutatjuk az `emBERT` modult, amely a `transformers` könyvtár segítségével lehetővé teszi kontextuális szóbeágyazás-alapú osztályozók integrálását az `e-magyar` rendszerbe. A modult főnévi csoport- és névelemfelismerésre tanítottuk fel. A modellek mindkét feladaton javítanak az eddigi legjobb eredményeken.
Kulcsszavak: BERT, `e-magyar`, névelem, chunking

1. Bevezetés

A gépi tanulásos nyelvi elemző rendszerek az utóbbi években drasztikus átalakuláson mentek keresztül. A hagyományos paradigma szerint minden szóhoz kézzel állítanak elő jellemzőket (*feature*). Ezek tipikusan nyelvi és írásképbeli jegyek, amiket általában nyelvészek hoznak létre. E jellemzők szolgálnak utána egy egyszerűbb, tipikusan off-the-shelf osztályozó (logisztikus regresszió, CRF) bemenetéül. A legtöbb szekvencia- vagy tokenklasszifikációs feladatot (szófajcímkézés, névelemfelismerés, szentimentelemzés) ilyen rendszerekkel oldották meg.

A mélytanulás elterjedésével a kézzel kiválasztott jellemzők fokozatosan a háttérbe szorultak. Helyüket a vektoriális szemantika világából ismert szóbeágyazás (*word embedding*) (Mikolov és mtsai, 2013; Pennington és mtsai, 2014) kezdte átvenni. Egy beágyazás minden szóhoz egy sokdimenziós, folytonos vektort rendel. Ezek a vektorok egy szemantikus teret feszítenek ki, ahol a hasonló jelentésű szavak vektorai egymáshoz közel esnek.

A beágyazások azonban nem csak szemantikai tartalommal bírnak, hanem implicit kódolják a szavak szintaktikus tulajdonságait is. Ez különösen alkalmassá teszi őket a gépi tanulók bemeneti jellemzőinek szerepére. A nagyobb szövegelemző láncok közül elsőként a Stanford CoreNLP szintaktikus elemzője egészült ki beágyazásokkal (Socher és mtsai, 2013). Mára a szóvektorok a legtöbb nyelvi elemző szoftverben megtalálhatóak.

A statikus beágyazások hátránya azonban, hogy egy szót minden környezetben ugyanaz a vektor reprezentál. Ez különösen a többjelentésű (pl. *körte*, *zebra*), vagy azonos alakú (pl. *dob*, *szív*) szavak esetén jelent problémát, mivel a szóvektor szükségszerűen a jelentések egyfajta amalgámja lesz, és nem fogja tükrözni a szó szintaktikai és szemantikai szerepét az aktuálisan elemzett mondatban.

A kontextualizált beágyazásoknál, mint az ELMo (Peters és mtsai, 2018) vagy a BERT (Devlin és mtsai, 2019), a szó vektora függ annak közvetlen környezetétől is. Ebből következik, hogy egy szó minden egyes előfordulásához más–más vektor tartozik. Ezek a vektorok implicit módon kódolják a szó szerepét a mondaton belül, teljesen kiváltva ezzel a kézileg összeállított feature-vektorokat. A beágyazásokat tipikusan nyelvmodellezéssel „*tanítják elő*".

Kontextuális beágyazáson alapuló rendszerek több nyelvi feladaton is felülmúlták hagyományos társaikat. A BERT, illetve követői, az XLNet (Yang és mtsai, 2019) és a RoBERTa (Liu és mtsai, 2019) főleg olyan, magasabb szintű feladatokban produkáltak erős eredményeket, mint a kérdésmegválaszolás, vagy GLUE (Wang és mtsai, 2018) teszt nyelvi megértést vizsgáló feladatai. Az ELMo és a Flair (Akbik és mtsai, 2018, 2019b) pedig névelemfelismerésben utasította maga mögé a korábbi rendszereket.

Ezek az eredmények a szövegelemző programokban is visszaköszönnek. A Flair rendszer[1] egy teljes nyelvi elemzőlánc, amelynek alapja a beágyazások szabad variálhatósága. Jelenleg ez nyújtja a legjobb teljesítményt névelemfelismerés mellett a főnévi csoport- és szófajcímkézésben is (Akbik és mtsai, 2019a).

A fenti eredmények természetesen angol nyelvre vonatkoznak. Ebben a cikkben megvizsgáljuk, hogy a kontextuális embeddingek képesek-e magyar nyelven is hasonlóan kimagasló teljesítményt nyújtani. Tesztfeladatnak a főnévi csoport- (*chunking*) és a névelemfelismerést (*named entity recognition, NER*) választottuk, mivel ezekre létezik angol precedens. Az elkészült modelleket egy új modulként integráljuk az `e-magyar` szövegelemző rendszerbe.

2. BERT

2.1. Miért a BERT?

Az előző fejezet végén felsorolt beágyazások közül a BERT-öt választottuk vizsgálatunk tárgyául. Ennek fő oka az, hogy a legtöbb beágyazás kizárólag angolul (esetleg kínaiul) elérhető. Tanításuk sok adatot és nagy számítási kapacitást igényel, ami a cikk írásakor nem állt rendelkezésünkre. A két kivétel az ELMo és a BERT, ahol elérhetőek előtanított többnyelvű modellek.

A kettő közül a BERT egyik előnye az ELMo-val szemben, hogy ún. *finomhangolásos* módszer (Devlin és mtsai, 2019): az előtanított modell könnyen finomhangolható a célfeladatra. Az ELMo ezzel szemben egy beágyazást ad, amit jellemzően feladatspecifikus architektúra bemenetén használnak. Mivel mi különálló modulban gondolkodtunk, meglévő rendszerek átalakítása nem jött szóba. A BERT másik előnye, hogy a magasabb szintű feladatokban jobb eredményeket ért el, mint az ELMo. A főnévi csoport- és névelemfelismerésre ez pont nem áll, ezért egy lehetséges további kutatási irány lehet az `emChunk` és `emNer` „ELMosítása".

[1] `https://github.com/zalandoresearch/flair`

2.2. A BERT bemutatása

A BERT egy többszintű, kétirányú Transformer kódoló (*encoder*) (Vaswani és mtsai, 2017). A modellt két nyelvmodellezési feladaton (Cloze teszt, következő mondat megjóslása) tanítják elő. A bemenetek a feladat jellegétől függően lehetnek mondatok, vagy mondatpárok. A szótár méretének korlátozása érdekében egy mondat nem szavak, hanem szóelemek (*wordpiece*) (Schuster és Nakajima, 2012) sorozata. A szótár a modellel együtt letölthető.

Az előtanított modellt minden célfeladathoz külön finomhangolják. Egy egyrétegű, előrecsatolt osztályozó hálót adnak hozzá, majd a BERTet és az osztályozót együtt tanítják.

Az angol BERT modellek két méretben hozzáférhetők: a `Base` modell 110 millió, a `Large` 340 millió paraméteres. A többnyelvű modell csak a kisebb, `Base` konfigurációban elérhető. Ezt 104 nyelvre tanították elő, és a szótára megközelítőleg 120 ezer szóelemet tartalmaz. A modellnek van nyers (`cased`) és kisbetűsített–ékezettelenített (`uncased`) változata is. Az e cikkben leírt kísérletek az előbbit használják, mivel egyrészt az angoltól eltérően a magyarban az ékezetek jelentésmegkülönböztető szereppel bírnak, másrészt névelemek azonosításakor fontos információ, hogy nagybetűvel kezdődik-e a szó.

2.3. Mennyire tud magyarul?

Mivel az általunk használt BERT 104 nyelven lett tanítva, felmerül a kérdés, hogy mennyire modellezi jól a magyar nyelvet. Kicsit pontosabban két kérdést fogalmazhatunk meg:

1. Mennyire tükrözik a szóelemek a magyar morfémákat?
2. Helyes szemantikai tartalommal bír-e egy-egy szóelem vektora, különös tekintettel a több nyelvben is előforduló homográf szóelemekre (pl. „*leg*”, „*old*”, stb.)?

Az első kérdés megválaszolásához szóelemekre bontottuk a Szeged NER korpusz összes szavát a többnyelvű modell tokenizálója, illetve egy több milliárd szavas magyar korpuszon tanított, 30 000[2] szavas BPE (Sennrich és mtsai, 2016) szótárral. Néhány kiragadott példát mutat be a 1. táblázat.

Mint látható, a szavak három csoportra oszthatók. Az első csoportba azok tartoznak, amiket a két tokenizáló hasonlóan kezel. Vagy azért, mert mindkét szótárban szerepelnek (és ezért maguk is szóelemek), vagy azért, mert egyik sem tudja értelmes egységekre bontani: utóbbira példa a „*zambiai*”.

A második csoport esetén a magyar szótár kevesebb, morfológiailag indokolt részre bontja a szavakat, míg a BERT szerinti tokenizálásban feltűnnek szemantika nélküli n-gramok is. Ennek megfelelően a többnyelvű változat mindig több szóelemből áll.

A harmadik csoportban az olló tovább nyílik: a magyar BPE tokenizálás változatlanul szemantikus, míg a BERT szóelemei véletlenszerű n-gramok. A

[2] Ez megegyezik az angol BERT szótárának méretével.

Szó	Többnyelvű	Magyar
Nemzeti	Nemzeti	Nemzeti
Andersen	Andersen	Andersen
labdarúgó	labdarúgó	labdarúgó
zambiai	za mbia i	z amb iai
megmaradt	meg maradt	megmaradt
hétfő	hét f ő	hétfő
keddtől	ke dd től	kedd től
edényben	ed ény ben	edény ben
Hétfőn	H ét f ő n	Hétfőn
tájékoztatták	tá j ék ozta tták	tájékoztat ták
leggazdagabb	leg ga zda gab b	leggazdagabb
elpártolt	el pá rto lt	el párt olt

1. táblázat. Néhány szó szóelemekre bontva a többnyelvű BERT szótára és egy magyar korpuszon épített BPE szótár alapján

hosszabb szavak lefedéséhez a többnyelvű tokenizálónak akár 4-5 szóelemere is szüksége van (a magyar BPE-nek elég 1-2). A „*hétfő*” és a „*Hétfő*” eltérő felbontása pedig arra utal, hogy a mondatkező szavak és névelemek szóelemekké tokenizálása különösen problémás lehet.

A fenti megfigyeléseket a 2. táblázat is megerősíti. A többnyelvű BERT átlagosan 50%-kal több szóelemet állít elő, mint a magyar BPE. A jelenség azonos mértékben érvényes csak a szótípusokat vagy a teljes korpuszt tekintve is. Mivel a leggyakoribb funkciószavak („*a*”, „*az*”, „*és*”) és írásjelek részei mindkét szótárnak, ez arra utal, hogy a gyakori szavak is konzisztensen rosszabb reprezentációt kapnak a többnyelvű BERTben.

Érdekes módon a nagybetűs szavak felbontásában nincs jelentős (kvantitatív) különbség a két szótár között: mindkét szótár átlagosan 4–5 szóelemre osztja őket. Ez a kisbetűs szavakhoz képesti relatív ritkaságukkal magyarázható, ugyanakkor előrevetíti, hogy a BERT (többnyelvű vagy sem) nem feltétlenül optimális névelemfelismerésre.

A második kérdés részletes megtárgyalása meghaladja e cikk kereteit. Implicit választ a két nyelvi feladaton elért eredmények adnak az 5. fejezetben.

3. Az `emBERT` modul

Fontos szempont volt, hogy az elkészült modelleket a kutatók, illetve nyelvfeldolgozás iránt érdeklődők számára egyszerűen hozzáférhetővé tegyük. E célból döntöttünk a modellek `e-magyar` rendszerbe (Váradi és mtsai, 2017) integrálása mellett. Az `e-magyar` új verziója, az `emtsv`[3] (Indig és mtsai, 2019) jelentősen

[3] `https://github.com/dlt-rilmta/emtsv`

Szóalak	Többnyelvű BERT	Magyar BPE	Különbség
kisbetű	2.24	1.34	67%
nagybetű	1.86	1.75	6%
együtt	2.14	1.44	49%
kisbetű (típus)	3.97	2.41	65%
nagybetű (típus)	4.65	4.27	9%
együtt	4.12	2.83	45%

2. táblázat. Átlagos szóelemszám szavanként / típusonként

megkönnyítette új modulok hozzáadását az elemzőlánchoz. Így született meg az `emBERT` modul.

Az `emBERT` követi az `emtsv` modulok konvencióit. Egyfelől működik önálló Python modulként, másfelől (opcionális) része az `e-magyar` elemzőláncnak. Telepítése után elérhetővé válnak a `bert-base-chunk`, `bert-max-chunk`, és `bert-ner` eszközök. Ezek tokenizált szöveget várnak bemenetükön, ezért az `emToken` futtatása előfeltétele a működésüknek. A többi, magasabb szintű `e-magyar` modultól (mint pl. az `emChunk` és az `emNer`) eltérően azonban az `emBERT` morfológiai információt nem igényel, ezért a morfológiai elemző és a lemmatizáló futtatása nem szükséges.

Mivel a BERT modellek (még `Base` konfigurációban is) nagyok, a modul nem tartalmazza őket. Ehelyett mind a három eszköz első meghívásakor letölti a saját modelljét az `emBERT_models` GitHub repozitóriumból[4].

A BERT finomhangolásához és futtatásához a HuggingFace `transformers`[5] (Wolf és mtsai, 2019) programkönyvtárat használtuk. A csomag előnye, hogy a BERT mellett tartalmazza más Transformer-alapú beágyazások (XLNet, RoBERTa) implementációit is. Ez lehetővé teszi később más beágyazások kipróbálását és integrálását a modulba.

A többi `e-magyar` modullal szemben az `emBERT` tartalmazza mind a tanító, mind a modelleket futtató kódot. Két okból választottuk ezt a megoldást: egyrészt a kód bonyolultsága nem indokolta a két funkció kettéválasztását; másrészt így a felhasználók egy kész csomagot kapnak, amivel kedvükre kísérletezhetnek. A kód a többi `e-magyar` modulhoz hasonlóan GitHubon[6] érhető el.

4. Kísérletek

A modellek képességeit két feladaton: főnévi csoport- és névelemfelismerésen mértük. A modelleket korábbi eredményekkel való összehasonlíthatóság érdeké-

[4] `https://github.com/DavidNemeskey/emBERT_models`
[5] `https://github.com/huggingface/transformers`
[6] `https://github.com/DavidNemeskey/emBERT`

ben a vonatkozó szakirodalomban használt korpuszokon tanítottuk és értékeltük ki.

A magyar statisztikai NP-felismerők (A `hunchunk` (Recski, 2010) és utódai) mindegyikét a Szeged Treebank 2.0 (Csendes és mtsai, 2005) korpuszon tanították. Mi is hasonlóképpen jártunk el: a 82 099 mondatos korpuszt korpuszt véletlenszerűen, 80%-10%-10% arányban osztottuk fel tanító-, validációs és teszthalmazokra. Mind a két alfeladatot (minimális és maximális főnévi csoportok) ugyanúgy futtatuk: az alap BERT modellt 4 epochon keresztül finomhangoltuk, majd kiértékeltük a teszthalmazon. A validációs halmaz alapján *early stopping*ra nem volt szükség.

A névelemfelismerőt a Szeged NER korpuszon (Szarvas és mtsai, 2006), a Szeged Treebank részhalmazán finomhangoltuk. Mivel a NER korpusz jóval kisebb, mint a teljes Treebank (a három vágás 8172–502–900 mondatos), ezért a modellt több, különböző konfigurációval is feltanítottuk. A legjobb modell 30 epochon keresztül tanult 10^{-5}-ről lineárisan csökkenő tanulási rátával.

A kísérletekhez a korábban említett `transformers` könyvtár PyTorch (Paszke és mtsai, 2017) verzióját használtuk. A tanítást párhuzamosan futtattuk 3 db GeForce RTX 2080 Ti kártyán, 16-os batch size-zal. Ezzel a konfigurációval mind a legjobb NER modell, mind a (jóval kevesebb epochig tanított) chunking modellek 3 óra alatt tanulnak fel. A chunkinghoz a hiperparaméterek többségét az alapértelmezett értéken hagytuk. A NER esetében több hiperparaméter-beállítást is kipróbáltunk, de végül (az epochszám és a tanulási ráta kivételével) itt is az alapértelmezett értékek bizonyultak a legjobbnak.

A tanítás pontos paraméterei a letöltött modellekhez tartozó konfigurációs file-okban megtekinthetők.

5. Eredmények

5.1. Főnévi csoportok

Az `emBERT` és a `hunchunk` család eredményeit a 3. táblázat foglalja össze. Mint látható, az `emBERT` mindkét korábbi rendszernél jobban teljesít, és mind a minimális, mind a maximális NP-k azonosításában state-of-the-art eredményt ér el.

A különbség minimális NP-k esetében nem jelentős; a maximális csoportokon elért F1 érték viszont szignifikánsan, másfél százalékkal jobb, mint az `e-magyar`-ban jelenleg (`emChunk` néven) működő HunTag3.

5.2. Névelemek

Névelemfelismerésben a kép vegyesebb (4. táblázat). Az `emBERT` jelentősen, 2%-al magasabb F1-et ér el, mint Szarvas és mtsai (2006) és Varga és Simon (2007), de a HunTag3 eredményétől elmarad. A spaCy az összehasonlítás szempontjából nem releváns, mivel a tanítóadata ki lett bővítve a hunNERwiki korpusszal (Nemeskey és Simon, 2012); kizárólag a teljesség kedvéért szerepel a táblázatban.

Rendszer	Minimális	Maximális
`hunchunk`/HunTag (Recski, 2010)	95,48%	89,11%
HunTag3 (Endrédy és Indig, 2015)	–	93,59%
`emBERT`	**95,58%**	**95,05%**

3. táblázat. A magyar főnévi csoport-felismerők összehasonlítása

Rendszer	F1
(Szarvas és mtsai, 2006)	94,77%
`hunner` (Varga és Simon, 2007)	95.06%
HunTag3 (Endrédy és Indig, 2015)	**97.87%**
`emBERT`	97,08%
spaCy[7]	*93,95%*

4. táblázat. A magyar névelemfelismerők összehasonlítása

A NER tanítása alatt belefutottunk abba a problémába, ami minden gépi, de különösen mélytanuló rendszer rákfenéje: az eredmények erősen függenek a tanítás hiperparamétereitől, a megfelelő hiperparaméterek megtalálása azonban extrém módon erőforrásigényes. A Szeged NER-hez hasonló, apró korpuszok esetén ez a hatás hatványozottan jelentkezik, mivel a modell nagyságrendekkel több paraméterrel rendelkezik, mint ahány tanítópélda rendelkezésre áll. A megoldás egy, a jelenleginél nagyobb NER korpusz (például a hunNERwiki egy ellenőrzött minőségű részhalmaza) lehetne.

6. További kutatás

Az `emBERT`, bár javít a korábbi legjobb eredményen NP-felismerésben, több szempontból is proof-of-conceptnek tekinthető. Az alábbiakban sorra vesszük ezen szempontokat, és a kapcsolódó lehetséges kutatási irányokat.

Egyrészt láttuk, hogy a többnyelvű BERT használata mindenképpen szuboptimális: mind a szóelemek, mind a teljes modell kénytelen a (viszonylag szűkös, hiszen csak `Base` változat) kapacitását 104 nyelv között megosztani. Egy magyar korpuszon feltanított BERT, különösen a `Large` modell, minden bizonnyal további javulást érne el. A jövőben tervezzük ilyen modellek tanítását és nyilvánosságra hozását.

Másrészt a BERT csak a jéghegy csúcsa; számos egyéb kontextuális szóbeágyazás létezik, mint az ELMo, a RoBERTa, vagy a Flair. Ahogy láttuk, ezek bizonyos feladatokban – pl. névelemfelismerésben is – felülmúlják a BERT-öt. Reményeink szerint ezen beágyazások magyar változata is elkészülhet, mely esetben természetesen integráljuk őket az `emBERT`-be.

Harmadrészt, a névszói csoport- és névelemfelismerés mellett érdemes lenne megvizsgálni más nyelvfeldolgozási lépések BERT-ösíthetőségét. A nyilvánvaló jelölt a morfológiai elemzés, amire már létezik mélytanulásos megoldás (Ugray, 2019). Emellett – a GLUE-hoz (Wang és mtsai, 2018) vagy SQuAD-hoz (Rajpurkar és mtsai, 2016) hasonló magyar nyelvi erőforrások megléte esetén – olyan, magasabb szintű feladatokra is adaptálni lehetne a modult, mint a szentimentelemzés, parafrázisok felismerése, vagy kérdésmegválaszolás. Ezzel pedig az `emBERT` a meglévő funkciók javításán felül új képességekkel is fel tudná ruházni az `e-magyar`t.

7. Összegzés

A cikkben bemutattuk az `e-magyar` szövegelemző rendszer egy új modulját. Az `emBERT` lehetővé teszi kontextuális szóbeágyazás-alapú osztályozók integrálását az `e-magyar`ba. A többnyelvű BERT modellt névszói csoport- és névelemfelismerésre tanítottuk fel. A modellek összemérhetőek az eddigi legjobb eredményekkel, vagy javítanak is rajtuk.

Az `emBERT` számos továbbfejlesztési lehetőséggel rendelkezik. A modul könnyen kiterjeszthető más mély beágyazások, illetve nyelvi feladatok támogatására, amennyiben a vonatkozó erőforrások (maga a beágyazás, tanítókorpusz) elérhetővé válnak.

Köszönönetnyilvánítás

A kutatást részben a 2018-1.2.1-NKP-2018-00008 *A mesterséges intelligencia matematikai alapjai* és az NKFIH 120145-ös *Szószerkezet felismerése mélytanulással* projektek támogatták. A finomhangolási kísérletek egy részét az NVIDIA által adományozott grafikus kártyákon futtattuk.

Hivatkozások

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019a), `https://www.aclweb.org/anthology/N19-4010`

Akbik, A., Bergmann, T., Vollgraf, R.: Pooled contextualized embeddings for named entity recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 724–728. Association for Computational Linguistics, Minneapolis, Minnesota (06 2019b), `https://www.aclweb.org/anthology/N19-1078`

Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649. Association for Computational Linguistics, Santa Fe, New Mexico, USA (08 2018), https://www.aclweb.org/anthology/C18-1139

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Lecture Notes in Computer Science: Text, Speech and Dialogue. pp. 123–131. Springer (2005)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of NAACL (2019)

Endrédy, I., Indig, B.: HunTag3, a General-purpose, Modular Sequential Tagger – Chunking Phrases in English and Maximal NPs and NER for Hungarian, p. 213–218. Uniwersytet im. Adama Mickiewicza w Poznaniu, Poznan (2015)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – Egy formátum mind felett [emtsv – One format to rule them all]. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2019)

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (szerk.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013), https://bit.ly/39HikH8

Nemeskey, D.M., Simon, E.: Automatically generated ne tagged corpora for english and hungarian. In: Proceedings of the 4th Named Entity Workshop. pp. 38–46. Association for Computational Linguistics (2012)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017)

Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics (2014), http://www.aclweb.org/anthology/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics (2018), http://aclweb.org/anthology/N18-1202

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–

2392. Association for Computational Linguistics, Austin, Texas (11 2016), https://www.aclweb.org/anthology/D16-1264

Recski, G.: Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel. In: Tanács, A., Vincze, V. (szerk.) VII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 333–341 (2010)

Schuster, M., Nakajima, K.: Japanese and korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5149–5152. IEEE (2012)

Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (8 2016), https://www.aclweb.org/anthology/P16-1162

Socher, R., Bauer, J., Manning, C.D., Andrew Y., N.: Parsing with compositional vector grammars. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). pp. 455–465. Association for Computational Linguistics, Sofia, Bulgaria (2013)

Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. In: Discovery Science, 9th International Conference, DS 2006, Barcelona, Spain, October 8-10, 2006, Proceedings. pp. 268–278 (2006)

Ugray, G.: Pos-tagging and lemmatization with a deep recurrent neural network. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2019). pp. 215–224. Szeged (2019)

Varga, D., Simon, E.: Hungarian named entity recognition with a maximum entropy approach. Acta Cybern. 18(2), 293–301 (Feb 2007)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: e-magyar: digitális nyelvfeldolgozó rendszer. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). Szeged (2017)

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multitask benchmark and analysis platform for natural language understanding (2018)

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing (2019)

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding (2019)

Szerzői index, névmutató

Ács, Judit, 171

Beke, András, 95
Berend, Gábor, 3, 43
Bilicki, Vilmos, 43
Bobály, Gábor, 261

Damsádi, Nóra, 103
Dömötör, Andrea, 385

Gosztolya, Gábor, 219, 313
Gráczi, Tekla Etelka, 103
Grósz, Tamás, 73
Gyimóthy, Tibor, 191

Horváth, Csilla, 261
Huszár, Anna, 103

Indig, Balázs, 29

Jenei, Attila Zoltán, 59

Kalivoda, Ágnes, 29
Kappel, Péter, 369
Kicsi, András, 15, 115, 191
Kiss, Gábor, 59, 83
Kiss, László, 333
Kmetty, Zoltán, 333
Kornai, András, 171
Kovács, Viktória, 129
Krepsz, Valéria, 103

Laki, László János, 155, 181, 343

Makrai, Márton, 273
Markó, Alexandra, 103
Mészáros, Evelin, 289
Mittelholcz, Iván, 29
Modrián-Horváth, Bernadett, 369

Nagy, Balázs, 333
Naszódi, Mátyás, 205
Nemeskey, Dávid Márk, 409
Németh, Péter, 15, 191
Nolda, Andreas, 369
Novák, Attila, 155, 303, 385
Novák, Borbála, 303

Pašić, Azra, 83
Perlaki, Attila, 343
Péter, Róbert, 43
Pintér, Ádám, 313
Pólya, Tibor, 323
Pusztai, Péter, 15, 115, 191

Ring, Orsolya, 333, 357

Sass, Bálint, 29, 399
Seres, József, 43
Simon, Eszter, 29
Szabó Ledenyi, Klaudia, 15, 191
Szabó, Endre, 115
Szabó, Martina Katalin, 333, 357
Szántó, Zsolt, 43
Száraz, Bettina, 103
Szaszák, György, 95, 245
Sztahó, Dávid, 83, 95

Tóth, László, 313
Trencsényi, Réka, 233
Tündik, Máté Ákos, 245

Vadász, Noémi, 29, 141
Vargáné Drewnowska, Ewa, 369
Vetráb, Mercedes, 219
Vidács, László, 15, 115, 191
Vincze, Veronika, 261, 333, 357

Yang, Zijian Győző, 155, 181, 343, 385