Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian

Bálint Döbrössy[1], Márton Makrai[2], Balázs Tarján[1], and György Szaszák[1]

[1]Dept of Telecommunications and Media Info, Budapest University of Technology and Econ
[2] Research Institute for Linguistics of the Hungarian Academy of Sciences
[1] balint.dobrossy@gmail.com, {tarjanb, szaszak}@tmit.bme.hu
[2]makrai.marton@nytud.mta.hu

Abstract

For morphologically rich languages, word embeddings provide less consistent semantic representations due to higher variance in word forms. Moreover, these languages often allow for less constrained word order, which further increases variance. For the highly agglutinative Hungarian, semantic accuracy of word embeddings measured on word analogy tasks drops by 50-75% compared to English. We observed that embeddings learn morphosyntax quite well instead.

Therefore, we explore and evaluate several sub-word unit based embedding strategies – character n-grams, lemmatization provided by an NLP-pipeline, and segments obtained in unsupervised learning (morfessor) – to boost semantic consistency in Hungarian word vectors. The effect of changing embedding dimension and context window size have also been considered. Morphological analysis based lemmatization was found to be the best strategy to improve embeddings' semantic accuracy, whereas adding character n-grams was found consistently counterproductive in this regard.

1 Introduction

Word embeddings show amazing capabilities in representing semantic relations, which has been demonstrated in analogical reasoning tasks (Mikolov et al., 2013b; Gladkova and Drozd, 2016). They are also capable of learning morphosyntax, showing again a consistent mapping of grammatical operations, i.e. inflections (see Section 2). Word embeddings obtain such semantic and syntactic capabilities by matching the words to their observed contexts (or vice versa). Since the size of the word vector table is the vocabulary size times the embedding dimension, for languages with rich morphology (especially agglutinative ones), this results in huge matrices (Takala, 2016). The vocabulary needs to be increased for morphologically rich languages to ensure a high enough coverage for the overall occurring words. Furthermore, to obtain a reliable estimate of word vectors, a larger training corpus is required so that theoretically the same convergence of the estimation can be reached than for a non agglutinative language. Finally, morphologically rich languages can express grammatical relations through suffixes (i.e. case endings) and hence let the word order becoming less constrained than in configurational languages. This can result in higher context variability, which translates again into less accurate estimates (i.e. the effect of migrating words outside the context window can be imagined as a kind of smoothing, making representation more blurred). Augmenting the size of the context window is not a effective counter-measure, as it will result again in higher variability of the context.

Bojanowski et al. (2017) proposes character level enhancement for word embeddings to overcome difficulties caused by unseen or rare words. It is demonstrated for a large set of languages that adding character n-grams to the embeddings can be a powerful way of generating word vectors for unseen words, and this augments both semantic and syntactic consistency (and accuracy) of the embeddings. However, Bojanowski et al. (2017) tests no highly agglutinative language for their embeddings' syntactic and semantic accuracies with and without n-grams.

We conduct proper evaluation on an analogy set for Hungarian (Makrai, 2015) designed according to the standard Mikolov et al. (2013a), and show that the already weak baseline semantic accuracy consistently decreases when character n-grams are added. On the other hand, embeddings learn the complex Hungarian morphosyntax quite

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 187–193
Florence, Italy, August 2, 2019.

well. Our ambition in this work is to address these issues emerging from large vocabulary and less constrained word order. We systematically investigate and analyze sub-word embedding strategies for the very highly agglutinating Hungarian language. We are basically interested in benchmarking syntactic and semantic accuracies with each of the methods, therefore we are primarily engaged in testing morphological analysis, lemmatization and stemming based alternatives.

2 Related work

The closest work to ours is a concurrent study (Zhu et al., 2019) of subword models especially for morphologically rich languages across different tasks. Unfortunately they miss Hungarian, which leaved a huge gap, as they find that performance is both language- and task-dependent. They find that unsupervised segmentation (e.g., BPE, Morfessor, see later in this section) is sometimes comparable to or even outperform supervised word segmentation.

Morphology in word embeddings The morphologically informed approach to compositionally gained word embedding vectors start with Lazaridou et al. (2013) and Luong et al. (2013), who train a Recursive Neural Network, which builds representations for morphologically complex words from their morphemes.

The work of Soricut and Och (2015) can be regarded as the unsupervised counterpart of Mikolov et al. (2013b)-style analogical questions. Soricut induces morphological relations as the systematic difference of embedding vectors in an unsupervised manner. They evaluate on word-similarity.

Relying on existing morphological resources, Cotterell et al. (2016) introduce a latent-variable morphological model that extrapolates vectors for unseen words, and smoothes those of observed words over several languages.

Cao and Rei (2016) introduce a joint model for unsupervised segmentation and weighted character-level composition. Cotterell et al. (2018) compute supervised models for the same two sub-tasks of morphological analysis, also induces a canonical form (i.e. models orthographic changes).

Language modeling and characters Morphologically compositional language modeling proper begins with Botha and Blunsom (2014)'s decoder in machine translation to morphologically rich languages, which is unsupervised with respect to morphological segmentation. Cotterell and Schütze (2015) augment the log-bilinear language model (LM) (Mnih and Hinton, 2007) with a multi-task objective for morphological tags along with the next word.

Character n-gram features proved to be powerful as the basis of Facebook's fastText classifier (Joulin et al., 2016). Subword units based on byte-pair encoding have been found to be particularly useful for machine translation (Sennrich et al., 2016), and even in models based on matrix factorization (Salle and Villavicencio, 2018).

Hungarian In their de-glutinatve method, Borbély et al. (2016) and Nemeskey (2017) split all inflectional prefixes into separate tokens for better morphological generalization. Nemeskey opts for supervised morphological knowledge because of linguistic interpretability. Lévai and Kornai (2019) analyze Hungarian word embedding vectors grouped by the morphological tag of the corresponding word. They investigate whether the coherence of these classes correlate with the specificity or the frequency of the tag.

3 Experiments

3.1 Corpus, segmentation, and embeddings

For training the word vector models, we rely on the fastText (Joulin et al., 2016) tool, which also allows for augmentation with character n-grams, if desired. We do not use stemming, but go instead for some more sophisticated analysis. As we explained, our primary goal is benchmarking the individual approaches.

For a true morphological analysis, we use the magyarlánc (Zsibrita et al., 2013) toolkit, which provides lemmatization in the form of a stem plus a suffix series, also decomposed into individual component morphemes. Although some disambiguation capability arises from sentence level part-of-speech tagging, magyarlánc may end up with several hypotheses for the morphological composition of the input word. Fortunately this happens rarely at the lemma level. If still, the shortest lemma is used.

For unsupervised pseudo-morphemic analysis, we use Morfessor (Virpioja et al., 2013). Morfessor has been used to provide subword unit tokens for Automatic Speech Recognition in heav-

Parameter	Value range
Frequency cut-off	5
Min length of char ngram	none or 3
Max length of char ngram	none or 6
Embedding dimension	100-200
Context window	5–25
Learning rate (α)	0.05
α update interval	100
Number of epochs	15
Negative sampling loss	yes
Negative samples	5
Pretraining	none

Table 1: Embedding vector trainer parameters.

ily agglutinative languages, with improved accuracy (Enarvi et al., 2017) over word based vocabularies and models. Morfessor is based on statistical machine learning. In order to reflect that the provided subword units are not true morphemes in the grammatical sense, they are called morfs.

The text corpus we use is a contemporary dump of Hungarian language web pages constructed for this paper, which covers mostly online newspapers in various fields from years 2014-2018. The corpus has over 70 M word tokens. Text normalization is performed with a Python script.

3.2 Analogical questions

Our approach is to train word embeddings in different scenarios and assess syntactic and semantic accuracies based on a Hungarian analogy test (Makrai, 2015) that has been constructed according to (Mikolov et al., 2013a). For the semantic accuracy, we use `country-capital` and `country-currency` pairs. For the syntactic accuracy we use `singular-plural` for nouns, `present-past tense` for verbs and `base vs comparative forms` for adjectives.

3.3 Fasttext settings

There are three main parameters which are controlled during the experiments: (i) whether we use character n-gram augmentation or not; (ii) the size of the context window; and (iii) the target dimension of the resulting embedding vectors. We preferred to preserve all other parameters of fastText at their default value. The most important of these parameters are summarized in Table 1.

3.4 Embedding strategies

Word vectors (W) This constitutes our baseline. A standard word embedding is trained with fastText, no prior stop word filtering is applied.

Lemma vectors (L) The magyarlánc toolkit is used for morphological analysis. Lemmas are identified and used as embedded entities. Note that whereas ambiguity on the entire morphological composition may arise, ambiguity affecting the lemma's surface form is rare. If this still occurs, the shortest form is used.

Morf vectors (M) Running Morfessor yields a morf based split-up. Morfs become the modeling unit (subword unit). As an alternative, using the **root (R)** yielded by Morfessor is evaluated as well. The word embedding is trained on the corpus with words divided into segments (as if they were separate words). During testing in analogical questions, query words are also spitted to segments, and their vectors are computed as the sum of the segments' vectors.

Vector dimension is changed between 100 and 200. We did not consider using higher dimensions to avoid making down-stream applications heavy.

More experimental details and related work can be found in a longer version of this paper, which appeared at Repl4NLP 2019. We will refer to the individual setups by specifying the unit out of {W, L, M, R} and the dimension, e.g. L200 will refer to lemma as unit and 200-dimensional embeddings.

4 Results

4.1 Extending the context window

As we pointed out in Section 1, using wider context may help in overcoming the difficulties resulting from the less constrained word order of Hungarian. A wider context window allows for capturing words further apart, but it may have an adverse effect as well, because the context becomes more noisy (variable). Relative data sparsity may also be a problem when a larger context is considered. So basically our research question related to the context of a word is that whether the benefits of capturing further apart words can be superior compared to the negative effect of increasing variance w.r.t the occurring context words.

It has been reported (Lebret and Collobert, 2015) that semantic analogical questions benefit

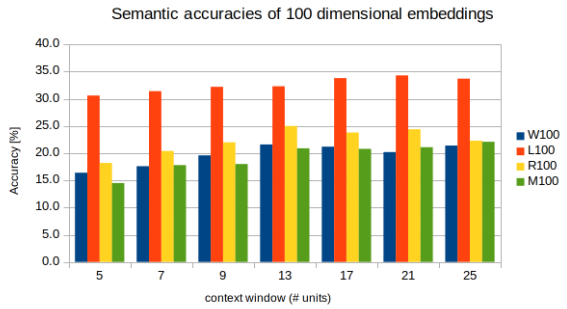

Figure 1: Semantic accuracies of Hungarian 100 dimensional embeddings with different strategies.

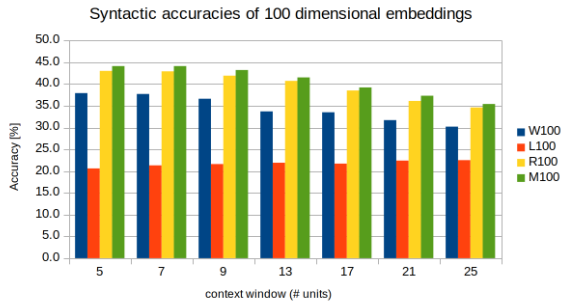

Figure 2: Syntactic accuracies of Hungarian 100 dimensional embeddings with different strategies.

from larger windows, while syntactic ones do not. On the contrary, experimenting with SVD models and different window sizes, Gladkova and Drozd (2016) find that all categories of analogical questions are best detected between window sizes 2–4, although a handful of them yield equally good performance in larger windows. They find no one-on-one correspondence between semantics and larger windows. We consider unusually large contexts of up to 25 words (see Table 1). i

Semantic and syntactic accuracies with 100 dimensional embeddings are shown in Figures 1 and 2, respectively. Comparing strategies, using the lemma (L) for embedding is yielding the highest semantic accuracy. Regarding the context window, our hypothesis that long context windows may be a better fit is confirmed. All the four strategies consistently show increasing semantic accuracy as context window is extended to cover 21 units. Compared to W, L embeddings yield higher semantic accuracy by 75%. Nevertheless, syntactic accuracies decrease tendentiously when extending the context window, which is a negative effect, most likely resulting from the higher variation seen in a larger window.

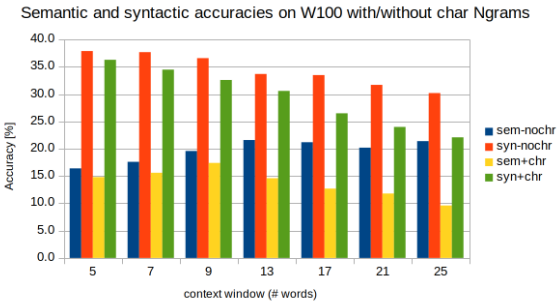

Figure 3: Semantic and syntactic accuracies of Hungarian 100 dimensional word embeddings with (chr) and without (nochr) character n-grams.

4.2 Adding character n-grams

We have already mentioned in the Introduction that in contrast to many other languages (Bojanowski et al., 2016), the very highly agglutinative Hungarian cannot profit from adding character n-grams to the embeddings: semantic (but also syntactic) accuracy gets lower. We suppose that this happens because agglutination is frequent and hence word vectors become universal (i.e. they cannot specialize for the context). The less constrained word order interplays in this, too.

Figure 3 shows how semantic and syntactic accuracies change when adding character n-grams (sem+chr and syn+chr, respectively) in the W100 case. We present again a trend with increasing context window size on the horizontal axis to allow for easy comparison with the previous results.

Regarding semantic accuracies, no benefit is registered when adding character n-grams with any of the 4 investigated embedding strategies.

Adding character n-grams becomes helpful at the syntax level in some cases, syntactic accuracies augment for the L100, L200 and R200 scenarios. Nevertheless, the basis is very low as for using the lemmas or morf roots, most of the morphosyntactic information is lost. Not surprisingly, semantics improves with a large window, while morphosyntax does not.

4.3 Embedding dimension

Figure 4 compares semantic accuracies of 100 and 200 dimensional scenarios with a context window of 21. Increasing the embedding dimension has a positive effect on semantic accuracies, as far as up to 50% relative increase in accuracy. Accuracy in individual relations (whose importance has been shown by Gladkova and Drozd (2016)) are

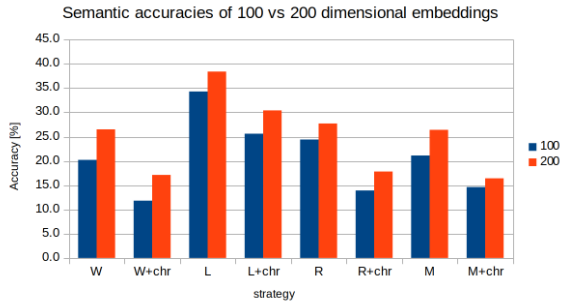

Figure 4: Semantic accuracies of Hungarian 100 and 200 dimensional embeddings with different strategies; context window covers 21 units.

capital-common-countries	66.0% (101/153)
capital-world	40.3% (2595/6441)
county-center	18.2% (12/66)
currency	6.4% (26/406)
family	16.5% (15/91)
Semantic	38.41% (2749/7157)

Table 2: Results in individual semantic relations with the best setting (magyarlánc, window 21, dimension 200, no character n-grams).

reported in Table 2. We can again observe that adding character n-grams consistently results in decreased semantic accuracy.

Increasing embedding dimensions above 200 could be expected to yield further improvement is semantic accuracies, but we did not address this issue in our current work, which focuses mostly on the modeling unit and its optimal context.

5 Conclusions

In this work, we analyzed embedding strategies for the morphologically very rich Hungarian language. Unlike may other languages, Hungarian cannot profit from character n-gram enhancement of word embeddings, whereas rich morphology results in very large vocabulary and less constrained word order, both contributing to very high variation in the data used for training the embeddings. Therefore we analyzed subword embedding strategies above the character level. Results showed that using the lemmas instead of the words was by far the most effective approach by maximizing semantic accuracy of the embeddings. Using the roots yielded by the `morfessor` tool also contributed to an increase in semantic accuracy, but to a smaller extent compared to lemmas learned in a supervised fashion. Obviously, syntactic accuracies were found decreasing when switching to lemma units. Adding character n-grams was counterproductive with any investigated strategy w.r.t semantic accuracy. Analyzing the effect of extending the context window showed that despite the higher variance of units seen in a larger context, embeddings can still profit from these to increase their semantic consistence. This founding was consistent with all investigated sub-word strategies, and is therefore an efficient way of dealing with the weakly constrained word order.

Future work may investigate whether results generalize to other embedding algorithms (besides fastText, the original and the enhanced (Mikolov et al., 2018) word2vec and the GloVe (Řehůřek and Sojka, 2010) implementations of the *continuous bag of words* and the *skip-gram* models could be tried); extend the ablation over dimensionality up to a few hundred dimensions; and analyze other morphologically rich languages (e.g. Finnish, Turkish, or Slavic languages). The bottleneck is that we are restricted to languages to which the analogical questions have been translated. As a reviewer noted, the semantic part of the Mikolov-style analogical questions consist of a handful of semantic relations between named entities. It is questionable how appropriate it is to use them for the evaluation of the embedding strategies, especially that of encoding lexical semantic relations and not the world knowledge. Gladkova and Drozd (2016) examine Mikolov et al. (2013b)-style analogical questions systematically, finding that different systems shine at different sub-categories of the morphological and semantic tasks. They publish a test set which is more difficult than existing ones. Translating this test set to morphologically rich languages would be very useful.

Acknowledgments

This work was supported by the Hungarian National Research, Development and Innovation Office under contract ID FK-124413: 'Enhancement of deep learning based semantic representations with acoustic-prosodic features for automatic spoken document summarization and retrieval'. Márton Makrai was partially supported by project found 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence and National Research, Development and Innovation Office grant #120145.

References

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. *arXiv preprint arXiv:1704.01938*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Piotr Bojanowski, Armand Joulin, and Tomas Mikolov. 2016. Alternative structures for character-level rnns. In *International Conference on Learning Representations, Workshop track (ICLR 2016)*.

Gábor Borbély, András Kornai, Dávid Nemeskey, and Marcus Kracht. 2016. Denoising composition in distributional semantics. In *DSALT: Distributional Semantics and Linguistic Theory*. Poster.

Jan A Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *ICML*, pages 1899–1907.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. In *Repl4NLP*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, Brussels, Belgium. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292.

Seppo Enarvi, Peter Smit, Sami Virpioja, Mikko Kurimo, Seppo Enarvi, Peter Smit, Sami Virpioja, and Mikko Kurimo. 2017. Automatic speech recognition with very large conversational finnish and estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(11):2085–2097.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proc. RepEval (this volume)*. ACL.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. ArXiv preprint arXiv:1607.01759.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526.

Rémi Lebret and Ronan Collobert. 2015. Rehabilitation of count-based models for word vector representations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 417–429. Springer.

Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113.

Dániel Lévai and András Kornai. 2019. The impact of inflection on word vectors. In *XV. Magyar Számítógépes Nyelvészeti Konferencia*.

Márton Makrai. 2015. Comparison of distributed language models on medium-resourced languages. In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*.

Tomas Mikolov, Kai Chen, G.s. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, volume 2013.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Language Resources and Evaluation Conference (LREC)*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM.

Dávid Márk Nemeskey. 2017. `emMorph` a hungarian language modeling baseline. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, pages 91–102, Szeged.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Alexandre Salle and Aline Villavicencio. 2018. Incorporating subword information into matrix factorization word embeddings. *arXiv preprint arXiv:1805.03710*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of NAACL*, pages 1627–1637, Denver, Colorado.

Pyry Takala. 2016. Word embeddings for morphologically rich languages. In *Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019. A systematic study of leveraging subword information for learning word representations. In *NAACL*. ArXiv:1904.07994.

János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc: A tool for morphological and dependency parsing of hungarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*, pages 763–771, Hissar, Bulgaria. INCOMA Ltd. Shoumen.