Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy

Radovan Garabík,
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia,
radovan.garabik@kassiopeia.juls.savba.sk
Denis Mitana,
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia,
denis.mitana@korpus.juls.savba.sk

Introduction

The Slovak language, as a "typical" Slavic language, belongs to the group of moderately inflected languages, with three or four genders, two grammatical numbers, all interacting with the inflections in somewhat complicated and unpredictable ways. The inflections are realized primarily by suffixes, but with many irregularities; one suffix encodes several relevant grammatical categories and the same suffix often reflects unrelated features in other words, a typical inflectional language not amenable to a heuristic analysis. Following these limitations, lemmatization is often an indispensable step in all kinds of text processing (starting with full-text search), and full morphosyntactic analysis or description (MSD) is the core of corpus linguistic research. Given the core importance of lemmatization and MSD in Slovak corpus linguistics, it is important to realize its limitations and recognize achievable accuracy. Since modern approaches aim to utilize deep learning and huge language models, we evaluate the accuracy of lemmatization + MSD in several common usage scenarios by comparing the state-of-the-art "classical" lemmatizer and MSD tagger *MorhoDiTa*, based on perceptron; and *spaCy*, using a multilingual BERT language model.

1. Dataset

Models were trained on a manually annotated corpus *r-mak-6.0* containing 720 documents, 77,671 sentences, 1,199,793 tokens, 55,090 unique lemmas, and 1,354 unique morphosyntactic tags. The composition of the corpus is 30.5% journalistic, 50.5% fiction, and 19.0% professional. To train the taggers, the corpus was divided randomly with stratification over documents into *train*, *dev*, and *test* segments in the ratio 8:1:1. The morphological database used for training MorphoDiTa contains 3,816,295 entries (i.e. distinct word-lemma-tag combinations), 114,634 unique lemmas, and 1,330,039 unique wordforms.

2. MorphoDiTa and SpaCy

MorphoDiTa [5] is a language-independent, open-source tool for morphological

analysis of natural language texts. MorphoDiTa has been extensively used in Slovak language corpora, and the tagger issued in the web interface provides access to lemmatization and tagging [4]. Part of MorphoDiTa is a statistical guesser for out-of-vocabulary (OOV) words, trained on suffixes. We use suffixes of at most 3 length, with 8 rules per suffix. To improve the accuracy of the guesser on real-world texts, we postprocess the guesser output and filter the list of possible lemmas to prefer tokens that appeared (as raw wordforms) in the corpora *prim-9.0-juls-sane* [1] and *Araneum Slovacum IV Maximum* [2]. We also include several heuristic rules to filter out implausible combinations of lemmas and tags and directly assign tags for numbers, punctuation, and other symbols. SpaCy is a language-independent, open-source, and production-ready Natural Language Processing (NLP) library. It comes with a wrapper package of the state-of-the-art Hugging Face's transformers architecture. For the Slovak language, spaCy officially supports only stop words and lexical attributes for general numerals. Although there is one online tool using spaCy available [6], it is not described further and is not available for unrestricted use. We use only MSD tagging and lemmatization components. As an architecture for the morphological analysis component, we use Transformer architecture based on the pre-trained multilingual BERT language model [3]. The lemmatization component is rule-based only. The rules applied are as follows: (1) if a given pair of word form and MSD tag is in the morphological database, then use the assigned lemma; (2) try to lemmatize a given pair of word form and MSD tag using the morphological suffix database. Postprocessing in the form of directly assigning tags for numbers, punctuation, and other symbols is the same as in MorphoDiTa.

3. Training and Evaluation

We summarise the accuracy of MorphoDiTa and spaCy output in the Table 1, where we take the *lemma+tag* accuracy to be the baseline. The "no OOV" row refers to MorphoDiTa accuracy calculated only on sentences containing only words known to the morphological database, thus describing a sort of "ideal" goal if the underlying morphological database has 100% coverage. As we can see, spaCy achieves higher accuracy in tagging. We suppose that more complex Transformer architecture handling a large number of output tags is better. On the other hand, spaCy is worse in lemmatization, apparently due to the rule-based approach. The most obvious and relevant single improvement of spaCy over MorphoDiTa is in disambiguating between singular masculine inanimate nominative and accusative (the word forms are identical), where apparently relatively free word order of Slovak requires better use of the context (or bigger context) to find out the correct case, where spaCy cuts the number of errors by two thirds compared to MorphoDiTa.

Conclusion

We calculated the accuracy of two state-of-the-art Slovak language lemmatizers and MSD taggers, one based on MorphoDiTa and the other one on spaCy. Over all, for

the combination of lemma+tag, spaCy with an accuracy of 95.6% overcame MorphoDiTa with 93.5%, with the most relevant single improvement in disambiguating between otherwise identical masculine inanimate singular nominative from the accusative, where better use of the context apparently helps to find out the correct case.

Table 1. Accuracy of various token annotations. CI means case insensitive.

Model name	Lemma	Lemma CI	MSD	POS	Lemma+tag	Lemma CI + tag
MorphoDiTa	98.24	98.95	94.19	98.06	93.50	94.03
spaCy	98.23	98.79	96.54	98.47	95.61	96.05
MorphoDiTa†	99.09	99.35	95.05	98.48	94.76	94.98

†no OOV

References

1. Slovenský národný korpus – prim-9.0-juls-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV. https://korpus.juls.savba.sk (2020), accessed: 2022-03-07

2. Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference* (pp. 257–264). Springer.

3. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). 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.

4. Garabík, R., & Bobeková, K. (2021). Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. *Slovenská reč*, *86*(1), 104–109.

5. Straková, J., Straka, M., & Hajič, J. (2014). Open-source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 13–18). Association for Computational Linguistics.

6. Wencel, M. Spracovanie prirodzeného jazyka – SpaCy Web App. https://spacy.tukekemt.xyz/analyze_sk (2021), accessed: 2022-03-07