

Variation in pre-modern Slavic corpus data and accuracy of neural tagging

Olga Lyashevskaya^{1,2}, Yves Scherrer³, Achim Rabus⁴

¹National Research University Higher School of Economics, ²Vinogradov Institute of the Russian Language RAS, ³University of Helsinki, ⁴Albert-Ludwigs-Universität Freiburg

Tagging pre-modern Orthodox Slavic varieties such as (Old) Church Slavonic, Old Russian or Middle Bulgarian poses several problems with respect to POS and full morphological tagging. Specifically, their rich morphology results in a large tagset, and both orthographic and morphological variation is abundant. The PROIEL and TOROT collections (Haug & Jøhndal 2008, Eckhoff & Berdicevskis 2015) provide large annotated corpora of various pre-modern Orthodox Slavic varieties. Earlier work has shown that these data can be used to train neural-network-based morphological taggers that achieve tagging accuracies over 90% (Scherrer et al. 2018).

In this paper, we assess the impact of variation on the tagging performance on the following aspects:

- (i) homogeneity in corpus annotation used for training;
- (ii) robustness of the tagger towards the diversity in language varieties in training and test collections;
- (iii) robustness of the tagger towards orthographic variation.

The mentioned training corpus collections are continuously expanded, while the Universal Dependencies annotation standard (Nivre et al. 2016) is being further harmonized. We investigate to what extent these developments enable us to train tagging models with broader coverage and better accuracy.

The term 'pre-modern Orthodox Slavic' subsumes several linguistic varieties with diverse characteristics, such as (Old) Church Slavonic, Old Russian, Middle Russian and Middle Bulgarian. Creating annotation tools for these varieties is a trade-off between specificity (independent models for each variety trained on small amounts of data) and universality (one single model for all varieties trained on large amounts of data). We run experiments with different data configurations and show how neural taggers fare with an intermediate approach that uses all available data and also exploits information about the varieties.

Orthographic variation is usually associated with data sparsity and leads to a greater rate of out-of-vocabulary items, words that have not been seen during the training phase but are present in the test data (OOV). While our earlier work only relied on character sequences, most recent work in Natural Language Processing relies on distributional representations of words (word embeddings). For OOV items, we create word embeddings by matching them with similar tokens that have been seen during training using vowel-sensitive Levenshtein distance. We discuss the challenges of creating and using such word embeddings in our setting. Whereas our approach favors the correct POS-tagging of orthographic variants, its drawback is the confusion of adverbs and adjectives and other words with overlapping paradigms.

Ultimately, our research shows that, when using neural taggers on pre-modern languages, tagging accuracy almost matches the accuracy of tagging modern high-resource languages with rich morphology.

References

Eckhoff, H. M. & Berdicevskis, A. (2015). Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15, pp. 9-25.

<https://torottreebank.github.io/>

Haug, D. T. T. & Jøhndal, M. L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Caroline Sporleder and Kiril Ribarov (eds.). *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27-34.

<https://proiel.github.io/>

Nivre, J. et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC 2016* (pp. 1659–1666). Portorož, Slovenia.

<http://universaldependencies.org/>

Scherrer, Y., Mocken, S., & Rabus, A. (2018). New developments in tagging pre-modern Orthodox Slavic texts. *Scripta & e-Scripta*, 18, 9–33.