DOUBLING THE AMOUNT OF TRAINING DATA: DOES IT HELP?

A New Training Corpus for Slovene and Its Impact on Automatic UD Annotation

Luka Terčon
Faculty of Computer and Information Science, University of Ljubljana

Nikola Ljubešić
Jožef Stefan Institute

Kaja Dobrovoljc
Faculty of Arts, University of Ljubljana

Relevant working groups: WG1, WG3

1.INTRODUCTION

The ssj500k training corpus was until recently the largest collection of manually annotated training data for Slovene (Krek et al., 2020), containing about 500,000 tokens, annotated on various levels of linguistic annotation. Recently, as part of the Development of Slovene in a Digital Environment project, we expanded this corpus with approximately 500,000 more tokens. The resulting dataset was named the Slovene Training Corpus (Slovene: Slovenski učni korpus) or the SUK corpus.

2.OBJECTIVE

The aim of the study was to describe how the new training data was used to train the CLASSLA-Stanza tool for automatic linguistic annotation (Ljubešić and Dobrovoljc, 2019). We specifically focused on investigating how varying the amount of training data impacts the performance of the tool for universal part-of-speech tag and universal dependency relation prediction. The impact of adding an inflectional lexicon to the classifier tool as a controlling element was also explored.

3.DOUBLING THE AMOUNT OF TRAINING DATA

In the first experiment (Table 1), a first set of models was trained on the ssj500k training data and a second set on the new SUK training data. This way we investigated how models trained on the original training set compare to models trained on the larger training set. While morphosyntactic and lemma annotations are present in the entirety of both the ssj500k and SUK corpora, UD syntactic dependency annotations are available for for only about 140,000 tokens in ssj500k and about 270,000 tokens in SUK. In effect, the amount of annotations on all three annotation layers were doubled. The results show a clear improvement in performance after doubling the amount of training data.

Annotation layer	Dataset	F1 Score
POS tagger	ssj500k	96.61
	SUK	**97.55**
Lemmatizer	ssj500k	98.89
	SUK	**99.33**
Dependency parser	ssj500k	87.78
	SUK	**91.06**

(top) Table 1: Comparison of model performance before and after doubling the amount of training data. Bold results are statistically significantly different to the alternative.

Annotation layer	Dataset	Lexicon usage	F1 Score
POS tagger	ssj500k	yes	**96.98**
		no	96.61
	SUK	yes	**97.94**
		no	97.55
Lemmatizer	ssj500k	yes	98.68
		no	**98.89**
	SUK	yes	99.11
		no	**99.33**
Dependency parser	ssj500k	yes	87.67
		no	87.78
	SUK	yes	91.11
		no	91.06

(right) Table 2: Comparison of model performance with and without lexicon usage. Bold results are statistically significantly different to the alternative.

4.ADDING AN INFLECTIONAL LEXICON INTO THE MIX

For the second experiment (Table 2), we analyzed model performance before and after adding an inflectional lexicon as a controlling element. This method restricts the model predictions to match the forms and combinations present within the lexicon. The lexicon clearly improves POS tagger performance, however on the level of lemmatization and dependency parsing the results are much more difficult to interpret. A subsequent error analysis showed that in some cases the lexicon guidance did improve the results, but also that the lexicon we used contains a number of automatically-generated entries, which proved detrimental to the performance of the annotation tool in some instances.

5.CONCLUSION

The experiments presented demonstrate that the increased amount of training data present in the new training corpus for Slovene improves the performance of tools for automatic grammatical annotation. This trend holds for all three inspected annotation layers. However, introducing an inflectional lexicon to limit the model predictions does not lead to a consistent improvement in the performance scores except for morphosyntactic tagging. This outcome reflects the great importance of good quality manually annotated data when it comes to both training corpora and inflectional lexicons.

RELATED LITERATURE

- Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Janez Brank. 2020. The ssj500k Training Corpus for Slovene Language Processing. pages 24–33.
- Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. pages 29–34. Association for Computational Linguistics

Univerza v Ljubljani Fakulteta za računalništvo in informatiko
Institut "Jožef Stefan" Ljubljana, Slovenija
Univerza v Ljubljani Filozofska fakulteta