A Dependency Parser for Maltese - Comparing the impact of transfer learning from Romance and Semitic Languages

Andrei Zammit, Slavomír Čéplö, Lonneke van der Plas, Claudia Borg

Tasks such as information retrieval, sentiment analysis and question answering require the processing of text analysis and natural language processing. Sentence parsing is one of the tasks performed in NLP to analyse the grammatical structure of a sentence, with the aim of determining the relationships between the words in a sentence.

This project builds upon previous work by Čéplö (2018) who built a dataset of dependency parses of 2,000 Maltese sentences, and Tiedemann and van der Plas (2016), who used cross-lingual transfer to bootstrap syntactic parsers for Maltese. We used machine learning techniques, namely deep learning, to build a computational model that could then be used to label new sentences. In order to train our models, we experimented with various language settings, including (i) training only on Maltese data, (ii) augmenting the training data with Spanish, Italian and English, and (iii) augmenting the training data further with either Hebrew or Arabic. The aim of the experiment was to determine the impact that the different languages can have on improving the parsing of Maltese sentences.

We obtained an 80.68% accuracy when using only Maltese data. Adding Spanish, Italian and English this increased to 89.77% but resulted in a slight drop when adding Arabic (88.86%) or Hebrew (88.61%).

The improvement in results when adding Spanish, Italian and English reflect the work by Tiedemann and van der Plas (2016), who also noted an increase in performance when using the same languages. However, the slight drop in performance when adding Arabic or Hebrew was unexpected. We posit that typology might be to blame for this result. The Arabic dataset used is taken from the UD treebank and uses Modern Standard Arabic, which basically reflects the same grammar as that of Sibawayh from circa 750 AD. This means that it is a completely different language typologically from Maltese today. Hebrew may share a few similarities with Maltese, but it has been argued that in its syntax, it is more Slavic and Germanic than Semitic. Tiedemann and van der Plas (2016) also confirmed that during their work they encountered problems with Arabic and they had to leave the language out of further experiments.

In future work, we plan to provide the dependency parser as an online service and create a framework so that newly annotated sentences can be easily corrected, and the dataset increased further.

References

Čéplö, S. (2018). Constituent order in Maltese: A quantitative analysis. PhD thesis, Charles University.

Tiedemann, J. and van der Plas, L. (2016). Bootstrapping a dependency parser for Maltese - a real-world test case. In *From Semantics to Dialectometry: Festschrift in honor of John Nerbonne*, pages 355–365, Milton Keynes, England. College Publications.