Creating, Evaluating and Enriching Parallel Corpora in the MaCoCu Project

We will outline part of the work done in the MaCoCu project: Massive collection and curation of mono-lingual and bi-lingual data. This project, which is funded by the Connecting Europe Facility, is a collaboration of 4 different partners: University of Groningen (us), Institut Jožef Stefan (Slovenia), University of Alicante (Spain) and Prompsit Language Engineering (Spain). A goal of this project is to release high-quality parallel corpora for under-resourced European Languages, by crawling top-level domains for each language. This first release contains data for Bulgarian, Croatian, Icelandic, Maltese, Slovenian and Turkish.

The first part of our work focuses on the evaluation of the crawled corpora. We aim to show that the crawled data is indeed of high quality by training Transformer-based Neural Machine Translation (NMT) systems on the parallel corpora. We compare a number of experimental settings: using just the MaCoCu data, and comparing adding the MaCoCU data to the largest currently available data sets to see if performance (still) improves. We evaluate performance across a number of languages, evaluation sets, domains and metrics and clearly find that the data is indeed of high-quality, with best performance for models that were at least partly trained on the MaCoCu data.

The second part of our work focuses on enriching the parallel corpora by automatically identifying which side of the parallel data is the original text, and which side is the translation. This information is often not present in parallel corpora, while it might be helpful to researchers, e.g. for those in the field of translation studies. Our approach consists of fine-tuning a multilingual pre-trained language model (XLM-RoBERTa) on parallel corpora from Europarl, one of the few that has this translation direction information already available. Additionally, we extracted training data ourselves by manually annotating a number of crawled domains for Croatian and Slovene. Our final sentence-level models perform quite well, with accuracies between 60 and 80%.