The CLIN2017 Shared Task

Tom Vanallemeersch
KU Leuven

Leen Sevens
KU Leuven

Our approach for translating 17th century to modern Dutch is fully data-driven and makes use of a parallel and a monolingual corpus. We train a statistical machine translation system from the parallel corpus, and derive a translation lexicon from pairs of words in the system that have high lexical probabilities and similar lengths and show strong Levenshtein similarity (we set weights for these three criteria through supervised machine learning). We also feed the parallel corpus to a module from our Text2Picto system for text to pictograph conversion, which aligns characters in sentences using Levenshtein, and learns rules for spelling variant generation from the aligned sentences. By combining the two resources derived from the parallel corpus, we generate potential sentence translations: we replace words found in the dictionary by their equivalent and generate spelling variants for the other words. Finally, we match parts of the potential translations to a suffix array built from a large corpus of modern Dutch, and validate strong matches. The suffix array is character-based, which allows us to match sequences of words and word fragments, and allows us to perform small adaptations between matching parts that are close to each other. We evaluated our approach using (a) the parallel 1637 and 1888 Bible versions provided for the shared task, (b) a 1-million sentence mixed-domain corpus of modern Dutch, and (c) test sentences extracted from "Het journaal van Bontekoe", of which the 17th century version and modern Dutch retranslation are available on dbnl.org. The evaluation shows that the spelling rules are very helpful, and that omitting the dictionary does not lead to a statistically significant difference in translation quality. However, the dictionary remains very useful as it strongly decreases computation time. Our evaluation also shows that performing machine translation as an initial step does not lead to improvement.