cSMTiser submission for the CLIN2017 shared task

Nikola Ljubešić
Jožef Stefan Institute, Ljubljana
University of Zagreb

Yves Scherrer
University of Geneva

Character-level statistical machine translation (CSMT) has shown to be useful not only for translating between closely related languages, but for the normalisation of various types of non-standard data such as dialects, historical texts and computer-mediated communication. Following the standard statistical machine translation architecture, a CSMT system is composed of a translation model - which is trained on parallel data - and of one or several language models - which are trained on data from the target language, in our case contemporary Dutch. An additional development corpus is required to estimate the respective weights of the different components of the system.

The translation model was trained using aligned sentence pairs from the different Bible versions provided by the organizers. In order not to 'pollute' the model with sentence pairs that look very different (either because of alignment errors or due to free translation), we removed the sentence pairs with a normalized Levenshtein similarity lower than 0.7. This filter yielded 87,644 sentence pairs coming from the (1637, 1888), (1637, 2010) and (1657, 1888) Bible combinations. We trained three 10-gram language models: the first one is based on the modern Bible sentences of the translation model, the second one uses the Dutch OpenSubtitles2016 corpus (http://opus.lingfil.uu.se/OpenSubtitles2016.php), and the third one uses the Dutch EUbookshop corpus (http://opus.lingfil.uu.se/EUbookshop.php). The development corpus consists of the Blankaart text together with its translation, made available by the organizers.

The system was trained using the Moses toolkit (http://www.statmt.org/moses/) and the cSMTiser scripts (https://github.com/clarinsi/csmtiser). As usual for normalization tasks, we disabled character reordering. The weights were tuned using the MERT algorithm and character error rate as a target metric. We found that most numbers were incorrectly normalized because the translation model barely contains any numbers. Therefore, we added a constraint that prevents tokens containing a numeral from being modified.

Note that one-to-many and many-to-one alignments (with '+' and '-' symbols) are not encoded in our submission. Also note that due to the character-based nature of our system, some tokens may not quite correspond to contemporary Dutch but still be closer to it than the unmodified original tokens. For this reason, we would be interested in an evaluation on character level, in addition to the token-level evaluation.