CLIN2017 Shared Task

Marijn Schraagen, Feike Dietz, Marjo van Koppen, Kalliopi Zervanou
Utrecht University

The approach is trained on the parallel text in the Statenvertaling (States Bible) from 1637 and 1888. The following steps are applied (in this order):

1. start with the alignment lexicon in the Shared Task baseline, obtained from aligned word pairs in same-length sentences.
2. perform alignment on unequal-length sentences using a custom algorithm, and extract additional translation pairs from this alignment.
3. apply manual translation rules, mostly morpho-syntactic modernizations.
4. extract multi-alignment pairs (1 word translated as 2 words or v.v.) from the parallel sentences and add these pairs to the translation lexicon.
5. apply manual phonetic and lexical rewriting rules.

An additional approach was tested using the Moses SMT toolkit, trained on the Statenvertaling. To this translation steps 3-5 were applied as above. For the Blankaart test file, the results are highly similar to the basic approach. Therefore, the Moses results are not included.

Both approaches are highly vocabulary-dependent. The new test set contains a rather different vocabulary, therefore the results of the modernization are not very accurate. On the Blankaart text the BLEU score is 0.46, only marginally higher than the original-translation score (0.43). In contrast, on the Statenvertaling test set the current method scores 0.65 (untranslated score: 0.13, baseline: 0.50).

Furthermore, the described approach suffers from 'overtranslation': several arguably correct results of the current approach are not present in the example translation of Blankaart, such as ofte-of, der-van de, hare-hun, 't-het, zo als-zoals, hebbe-heb, mate-maat, etc.

Translation pairs using POS tags has been implemented as well (i.e., 'sijne'+N > zijn or LID+'sijne' > zijne). This improves results on the Statenvertaling, but not on the current test set, and is therefore excluded.