The CLIN2017 Shared Task

Remko Boschker
University of Groningen

Rob van der Goot
University of Groningen

This system first generates a list of candidates for each word, and then uses the viterbi algorithm to rank these candidates. For each word candidates are generated by checking if the word does not contain any lower case letters, is in a small set of rules, is in a lexicon based on the Bible and example texts or if the parallel input text indicates that a word is in a dictionary. If either of these is the case, then a single possible translation is returned with an emission probability of one. If it is not, then all words in the Aspel dictionary that are within a Levenshtein distance of the length of the word divided by three and rounded up to the nearest integer, are considered as possible translations. Each of these possible translations is assigned an emission probability that is equal to its unigram count in a ngram model based on the Sonar corpus divided by the Levenshtein distance and a constant. If a translation is not in the unigram model, then it is assigned a count of 0.5. The emission probabilities of the possible translations are normalized to a sum of one. The transition probabilities between all possible translations of adjoining words is modeled by taking the bigram count from the ngram model and dividing that by the unigram count of the first word. If a bigram is not in the model it receives a count of 0.5. The transition probabilities between the possible translations of two adjoining words are normalised to a sum of one.