Marcel Bollmann, Stefanie Dipper, Florian Petran

Ruhr-Universität Bochum, Germany
(lastname)@linguistics.rub.de

We treat the task of translating 17th century Dutch into modern Dutch as a spelling normalization problem, and offer two different methods for performing the normalization: (1) the Norma tool (Bollmann, 2012), which implements a lexical mapping, a rule-based algorithm, and a distance-based algorithm; and (2) a character-based encoder/decoder neural network similar to Sutskever et al. (2014), consisting of a single bi-directional long short-term memory unit (LSTM) for the encoder, and an attentional LSTM for the decoder.

Both methods operate on wordforms in isolation, i.e. they do not take word context into account. Both are trained only on the 1637 to 1888 bible translation task; to produce the word-aligned training data, we align the bible versions using MGIZA,[1] resolving 1:n alignments using the underscore notation (`word_word`). During prediction, a lexical filter is used to restrict the wordforms that can be generated; for this, we use a combination of the 1888 bible and the Dutch part of CELEX (Baayen et al., 1995).

For preprocessing, we lowercase all tokens and remove all punctuation. Numbers and abbreviations are excluded from the normalization step, and simply copied verbatim. After normalization, punctuation is restored from the input file (i.e., no attempt is made to modernize it), while capitalization is restored using only a simple heuristic (of upper-casing sentence beginnings as well as single-letter abbreviations).

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2) (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bollmann, M. (2012). (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2).* Lisbon, Portugal.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (nips 2014)* (pp. 3104–3112).

[1] `https://github.com/moses-smt/mgiza`