PRHLT's submission to CLIN27 Shared Task: Translating Historical Text

Miguel Domingo, Francisco Casacuberta

Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València
Camino de Vera s/n, 46022 Valencia, Spain

midobal@prhlt.upv.es, fcn@prhlt.upv.es

Abstract

The PRHLT's approach to the shared task has relied on Statistical Machine Translation (SMT). Considering 17th century Dutch as a source language, and 21st century Dutch as the target language, we have trained an SMT system to perform such translation. As a training data, we've made use of the 17th to 19th century version of the *Bible* which was provided at the shared task. Additionally, we've used all the documents ranging from 17th to 21st century available at the *Digitale Bibliotheek voor de Nederlandse letteren*[1] to enrich the language model. The translation system was trained with the SMT state-of-the-art `Moses` toolkit (Koehn et al., 2007), using the MERT procedure (Och, 2003) for optimizing the weights of the log-lineal model, and estimating a 5-gram language model—using the improved KneserNey smoothing (Chen and Goodman, 1996)—with the SRILM toolkit (Stolcke, 2002).

References

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 310–318.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

[1] http://dbnl.nl/

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 257–286.