Neural Machine Translation for Literary Texts

Henny Sluyter-Gäthje
Universität Hamburg
xsluyter@informatik.uni-hamburg.de

Fabian Barteld
Universität Hamburg
fabian.barteld@uni-hamburg.de

Heike Zinsmeister
Universität Hamburg
heike.zinsmeister@uni-hamburg.de

Tag
Datum
Zeit
Raum

Rich vocabulary, variable domain (including direct speech) and a larger set of syntactic constructions are specific to literary text (van Cranenburgh & Bod 2017). These features should be preserved within translation as they form the reading experience. Statistical Machine Translation (SMT) approaches performed poorly on these specific features (e.g. Toral & Way 2015). On standard data, Neural Machine Translation (NMT) outperforms SMT systems in terms of fluency, syntax, word choice and the handling of rare words (e.g. Koehn & Knowles 2017). We present first experiments of NMT for literary texts comparing the performance of the Edinburgh NMT system (Sennrich et al. 2016) with the SMT system Moses (Koehn et al. 2007). The systems were trained with (i) in-domain data only, or (ii) in-domain data complemented with a larger set of out-of-domain data. For the SMT system we added monolingual data to generate the language model. We found that out-of-domain data did not bring any improvements and that the NMT system could not outperform the SMT system. In contrast to results in related work, the sentence length did not affect the score of the NMT system. Our manual analysis also showed (based on Popović et al. 2013) that the NMT system could not produce a better word choice which results in a fluent but often meaningless output. Confirming previous results, the SMT system often produced a disfluent output–due to more morphological and syntactical errors, but it made better lexical choices. To improve the results, we need a larger parallel literary dataset.

References: • Cranenburgh, A. v. & R. Bod. 2017. A Data-Oriented Model of Literary Language. In Proceedings of EACL 2017. • Koehn, P. & R. Knowles. 2017. Six challenges for neural machine translation. In Proceedings of the 1st Workshop on Neural Machine Translation. • Koehn, P. et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th ACL 2007. • Popović, M. et al. 2013. Learning from human judgments of machine translation output. In Proceedings of MT Summit. • Sennrich, R., B. Haddow & A. Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In Proceedings of the 1st Conference on Machine Translation. • Toral, A. & A. Way. 2015. Translating Literary Text between Related Languages Using SMT. In Proceedings of CLfL@NAACL-HLT.