NLP for German CMC Data

Thomas Proisl
Philipp Heinrich

We evaluate different approaches to standard low-level natural language processing (NLP) tasks on German natural language data from computer-mediated communication (CMC), in particular tokenization, part-of-speech tagging, and lemmatization. CMC data pose particular challenges to NLP systems because of many non-standard phenomena (e.g. colloquialisms, creative spellings, emojis, …). Whereas rule-based systems can solve the first step in the pipeline, tokenization (cf. Proisl and Uhrig 2016), with F1-scores above 99%, the steps further down the pipeline – POS tagging and lemmatization – are usually tackled with the help of machine learning (ML) methods. For POS tagging, we show that simpler ML methods such as linear classifiers are superior to deep neural networks when only a small amount of training data is available. Traditional POS taggers also benefit from additional resources such as Brown clusters or lexica (cf. Proisl 2018). For greater amounts of training data, or when domain adaptation is an option, deep neural networks outperform simpler approaches. We also present preliminary results on lemmatization, comparing different strategies on a newly created gold standard of roughly 23,000 tokens of running text.

References

Proisl, Thomas. 2018. “SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari et al., 665–670. Miyazaki: ELRA. http://www.lrec-conf.org/proceedings/lrec2018/pdf/49.pdf.

Proisl, Thomas, and Peter Uhrig. 2016. “SoMaJo: State-of-the-Art Tokenization for German Web and Social Media Texts.” In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, edited by Paul Cook at al., 57–62. Berlin: ACL. http://aclweb.org/anthology/W16-2607.