Presentation

Tatyana Ruzsics, Massimo Lusetti, Anne Göhring, Tanja Samardzic, and Elisabeth Stark

Neural Text Normalization with Adapted Decoding and POS Features

Abstract:
The task of text normalization aims at bringing non-canonical language, coming from speech and social media, to a standardized writing. This task is especially important for languages such as Swiss German, with strong regional variation and no written standard.
In this work, we propose a novel solution for normalizing Swiss German WhatsApp messages using the encoder-decoder neural machine translation (NMT) framework. We enhance the performance of a plain character-level NMT model with the integration of a word-level language model and linguistic features (POS tags). The two components are intended to improve the performance by addressing two specific issues. The former targets the fluency of the sequences predicted by NMT: it corrects a sequence which is not a proper word, despite being a likely sequence of characters. In addition, this modification targets the frequent cases where a contracted form corresponds to multiple normalized words, e.g. the word ‘kömmer’ ('we can') is mapped to the normalization form ‘können wir’. The latter component, the addition of POS tags, aims at resolving cases of word-level ambiguity. For example, the ambiguous input word Lüüt can be normalized as the noun Leute 'people' or läuten 'to ring' when used as a verb.
Our systematic comparison shows that the proposed model improves over the best previous solution. A thorough analysis of the compared systems' output shows that our two components produce indeed the intended, complementary improvements.

Intended audience:
Our intended audience includes both researchers and professionals interested in applying our resources in their own work.

Biography: Massimo Lusetti obtained a Master's degree in Multilingual Text Analysis at the University of Zurich, where he currently works as assistant. His main interests in NLP are machine translation and distributional semantics.

Organization: University of Zurich

Contact: tatiana.ruzsics@uzh.ch, massimo.lusetti@uzh.ch, goehring@cl.uzh.ch, tanja.samardzic@uzh.ch, estark@rom.uzh.ch