Mgr Roman Grundkiewicz

Poznań, 12.09.2017 r.

Streszczenie rozprawy doktorskiej pt.
„Algorithms for automatic grammatical error correction"
w języku angielskim

This thesis explores the problem of automated grammatical error correction (GEC) in texts written by non-native English speakers. Our main focus is the machine translation approach to GEC. To overcome the data sparsity problem, we have developed a method for the automatic extraction of potential errors from Wikipedia text edition histories, and created the largest publicly available error annotated corpus so far. We investigate the usefulness of automatic GEC-specific metrics on the basis of their correlation with human judgements by conducting the first large-scale human evaluation study of automated GEC systems. Our proposed phrase-based statistical machine translation (SMT) system achieved new state-of-the-art results on the CoNLL-2014 test data – a standard benchmark for GEC provided during the Conference on Natural Language Learning shared task in 2014. We have shown that parameter optimization towards the task-specific evaluation metric and new GEC-adapted dense features are crucial for building a reliable and effective SMT-based GEC system. We also examined two methods which incorporate discriminative components into the generative SMT log-linear model. In the case of the second method – the first reported application of sparse features to GEC – our results significantly improve over the previous state-of-the-art in the field.

Roman Grundkiewicz