Improving SMT-based Synonym Extraction across Word Classes by Distributional Reranking of Synonyms and Hypernyms

M. Bräuninger[1], M. Weller-Di Marco[2], S. Schulte im Walde[1]
[1] *Universität Stuttgart,* [2] *Universiteit van Amsterdam*
{maximilian.braeuninger,marion.di.marco,schulte}@ims.uni-stuttgart.de

Tag
Datum
Zeit
Raum

Finding synonyms is an important task in Natural Language Processing (NLP), e.g., for the creation of thesauri (Lin et al., 2003), and in performing and evaluating automatic machine translation (Lavie and Denkowski, 2009). Our work aims to extract synonyms by making use of statistical machine translation (SMT) methods relying on multilingual parallel corpora, as first described by Bannard and Callison-Burch (2005). In a first step, word alignments obtained from a parallel corpus are used in order to translate German words into English pivots, which are then re-translated into German. Using this method, a number of synonym candidates is created and ranked according to a combination of translation and re-translation probabilities.

In a second step two distributional semantic measures are introduced in order to re-rank the synonym candidates. The first measure based on feature overlap (Weeds and Weir, 2003) tries to identify hypernymy between the targets and the synonym candidates, and ranks hypernyms lower in the candidate list. The second measure relies on the distributional similarity between the targets and the candidates, ranking semantically highly similar candidates higher in the list, as done previously by Wittmann et al. (2014).

A gold standard across word classes (nouns, verbs, adjectives) is created using the online synonym section of the German dictionary *DUDEN*. While none of the re-ranking methods significantly outperforms the original approach, a number of interesting obeservations are found. For example, a manual evaluation suggested that approx. 20% of the invalid synonyms could be regarded as valid.

References: • D. Lin et al. (2003): Identifying Synonyms Among Distributionally Similar Words. In: *Proc. of Artificial Intelligence.* • A Lavie & M.J. Denkowski (2009): The METEOR Metric for Automatic Evaluation of Machine Translation. *Maschine Translation*, 23(2). • C. Bannard & C. Callison-Burch (2005): Paraphrasing with Bilingual Parallel Corpora. In: *Proc. of ACL.* • J. Weeds & D. Weir (2003): A General Framework for Distributional Similarity. In: *Proc. of EMNLP.* • M. Wittmann et al. (2014): Automatic Extraction of Synonyms for German Particle Verbs from Parallel Data with Distributional Similarity as a Re-Ranking Feature. In: *Proc. of LREC.*