Cross-linguistic differences in discourse marking: A case study of German-English texts

Frances Yung, Merel Scholman, Vera Demberg
Universität des Saarlandes
frances, m.c.j.scholma, vera @coli.uni-saarland.de

We present results on the cross-lingual alignment and annotation of discourse relations (DRs) in English-German parallel texts from the Europarl corpus, which is part of our project to create large parallel discourse-annotated corpora and lexicons in various languages.

The DiscoGeM Corpus (Scholman et al., 2022) includes crowdsourced annotations of about 400 DRs in English texts that were translated to German and 700 DRs in English texts translated from German, extracting from 15 and 18 documents respectively. We aim at creating a PDTB3-styled discourse annotated dataset with these documents where all explicit and inter-sentential implicit DRs are annotated and cross-lingually aligned. The process involves below steps:

1. English and German explicit DRs are identified together with the argument spans and classified into different sense types using the shallow discourse parsers (English: Knaebel 2021, German: Bourgonje, 2021).
2. The identified connectives are aligned using a combination of neural and statistical word alignment models (Östling and Tiedemann 2016, Dou et al. 2021). A focus is on the null alignments produced by the aligners, which reveals cases of implication and explication.
3. Implicit German DRs (consecutive sentences that are not connected by an explicit connective), and implicit English DRs not annotated in the DiscoGeM, will be annotated in two-step crowdsourcing annotation methodology (Yung et al., 2019)

The resulting dataset will allow us to extract a distributional bilingual connective lexicon for English and German, and provide data for the study of implication and explication of DRs in translation.

References: • Bourgonje (2021) Shallow discourse parsing for German. Phd-Thesis. • Dou & Neubig (2021) Word Alignment by Fine-tuning Embeddings on Parallel Corpora. EACL • Knaebel (2021) Discopy: A neural system for shallow discourse parsing. CODI • Östling & Tiedemann (2016) Efficient word alignment with markov chain monte carlo. The Prague Bulletin of Mathematical Linguistics • Scholman et al. (2022) Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. LREC • Yung et. al. (2019) Crowdsourcing discourse relation annotations by a two-step connective insertion task. LAW