Report on the Visit

Yannick Versley

March 31, 2015

Purpose of the STSM The broader purpose of the STSM was to study ways in which linguistic annotations of discourse-relational devices (DRDs) in one language can be used for annotation projection, in particular from English to other languages such as German and French, and how to exploit the result the results of annotation projection to yield (i) a list of DRDs in the target language, but also (ii) an initial classification into coarse categories of DRDs such as (for example) causal, adversative, or conditional connectives.

These goals, in particular (i), would be reached based on earlier research by Versley (2010) as well as Laali and Khosseim (2014), where Versley presents an approach that is purely based on word alignments whereas Laali and Khosseim present an approach that uses filtering by part-of-speech (POS) and syntactic information; While Versley's work is language-independent in design, but has not been evaluated intrinsically (Versley evaluates the approach as an end-to-end method to port a connective tagger between languages), Laali and Khosseim's POS and syntax filters are specific to French, therefore a declared goal of the STSM was to explore ways of having language-portable constraints expressing the same (or similar) information.

Description of the work carried out For the investigation into annotation projection, two datasets were for the language pairs EN-DE and EN-FR were taken into account, one being the EuroParl corpus in the most recent version from the OPUS collection, for which YV created word alignments using the POSTCat tool in advance of the STSM. The second dataset, which was used in preprocessing only, was the IWSLT14 corpus containing TED talks (i.e., spoken language).

For the EuroParl corpus, parses on the English and German sides already existed, whereas the French side in OPUS has a POS layer that is not compatible with existing resources (Achim Stein's TreeTagger model). The French side of the EuroParl data was parsed with the Bonsai constituent parser, which uses Petrov's PCFG-LA parser together with word clusters and morphological preprocessing, yielding tags and constituents according to the French Treebank (FTB) tagset.

For the English and French side of the IWSLT14 corpus, a new toolchain for preprocessing based on the ExportXML (EXML) format that allows to add further annotation layers beyond syntax – in this case, coreference annotations from the Stanford Sieve coreference resolver. For this corpus, the first step was a conversion from the DiscoMT shared task format (which contains, for each sentence-aligned segment, source and target tokens as well as word alignments), to ExportXML, which necessitates the insertion of sentence boundaries. The second step was carried out using ExmlPipe, a wrapper around different tools (Mate parser, PCFG-LA, CoreNLP) with the purpose of processing either raw text or EXML-format corpora into an EXML file with suitable annotation layers. For the English side, the CoreNLP pipeline (including POS/lemma/constitutents/dependencies/NER and coreference) was used, for the French side, a PCFG-LA model trained on the SPMRL'13 shared task version of the French Treebank was used. In a third step, the layers of the EXML files are converted into the data used by PyCWB (i.e., one file with sentence and document boundaries as well as all word-level annotations for the CQP index as well as files with syntactic annotations and word alignments).

The preprocessed EuroParl version was then the starting point for the extraction of candidates according to Versley (any string – continuous or not – that is aligned with a single- or two-part connective that has been identified on the English side), as well as according to Laali and Khosseim (any unigram or bigram that overlaps with a word aligned to whole or part of an identified English connective).

In a second step of candidate extraction, both the statistics used by Versley (average overlap) and those used by Laali and Khosseim (likelihood ratio test for association with the most typical relation, POS sequence) are extracted.

Main results obtained Among the main results are: firstly, a preprocessing pipeline based on the ExportXML format that allows to preprocess parallel corpora using different tools, including named entity recognition and coreference, which was not possible with the older system.

Secondly, detailed statistics on the distribution of POS sequences among the connectives tagged according to either of PDTB, HdK, and LexConn, yielding a picture of the principal differences of languages and/or schemes (for example, French sentence subordinators yield a `ADP CONJ` pattern for connectives like "après que", which would be atypical for German or English, but also `ADP DET NOUN CONJ` for connectives such as "*par le fait que*", which have no exact counterpart in German but at least structurally similar items such as "*unter der Voraussetzung (,dass)*" where, however, the process for preparing the German HdK dataset omits the sentence subordinator *dass*).

Thirdly, a new process for extracting candidates according to either Versley's or Laali and Khosseim's methods, including features defined by either work, which will allow a more direct comparison between the two approaches as well as a unified approach based on universal part of speech tags.

Future collaboration with host institution Besides the work based on the OPUS corpus collection, YV's stay in Uppsala has helped to deepen the existing and ongoing cooperation within the DiscoMT shared task, coorganized among others by Christian Hardmeier and Jörg Tiedemann from Uppsala and Yannick Versley from Heidelberg, with involvements from others, which has enabled fruitful discussions and actual progress on some data preparation as well as the baseline system for the Pronoun Prediction Task.

Planned publications Yannick Versley and Jörg Tiedemann plan to disseminate the results of the STSM in more elaborated form in a paper on connective projection.