TextLink: Structuring Discourse in Multilingual Europe

STSM Report

COST STSM Reference Number: COST-STSM-IS1312-34130
Period: 2016-06-20 to 2016-07-06
COST Action: IS1312
STSM type: Regular (from United Kingdom to Portugal)
STSM Applicant: Mr Samuel Gibbon, University of Edinburgh, Edinburgh (UK), samuel.gibbon@gmail.com
STSM Topic: Annotating Discourse Relations in TED Talks
Host: Amalia Mendes, University of Lisbon, Lisbon (PT), amalia.mendes@clul.ul.pt

As explained in my letter of motivation, the central goal of this STSM was to make a start on producing a corpus of parallel translated texts annotated with both implicit and explicit discourse relations. We were interested to see if the well-established annotation scheme used in PDTB could be applied to a different genre of text – prepared speech. We used the annotation scheme from PDTB 3 (sense relation taxonomy as of 28 May 2016) to annotate several English language TED Talks that had been translated into Portuguese, Turkish, Polish, German, and Russian. Other people participating in this exercise were Deniz Zeyrek (METU), Yulia Grishina (Univ of Potsdam), and Maciej Ogrodniczuk (IPIPam, Warsaw).

Our annotation and adjudication sessions in Lisbon proceeded in the following manner: First we individually annotated a given talk in our respective language, and then projected the English talk onto a large screen and went on to discuss each token, moving on when we reached agreement. Although this was a time-consuming process, especially at first, it resulted in a high level of cross language alignment with regards to the relation sense (e.g. temporal, concession, cause) and argument span. A number of interesting points and technical issues emerged from this process, of which the main ones are documented below.

Firstly, it is prudent to say that TED Talks are an excellent linguistic resource, given that a talk has often been translated into up to 150 different languages. Furthermore, they are not extremely technical and are generally interesting, thereby lessening the possibility of either annotator incomprehension or boredom. (For similar discussion, see Hovy & Lavid, 2010). However some issues arise when considering the processes of both transcription and translation.

Speech is continuous, and often one sentence flows into the next. One job of the transcriber is to use punctuation to parse each sentence, clause, phrase, and indeed each utterance into meaningful and readable text. Sometimes this is obvious, e.g. syntactically, or when the speaker takes a long pause between utterances, but other times less so. This may lead to differences in sentence tokenisation in different languages, since the translators listen to (and are hence influenced by) the talks as well by the English transcript. For example compare the following token in its English transcript and its Portuguese translation:

Talk 1927 (EN) *Environment includes energy consumption, water availability, waste and pollution, just making efficient uses of resources.*

Talk 1927 (PO) *Ambiente inclui consumo de energia , disponibilidade de água , lixo e poluição . Trata de_ o uso eficaz de_ os recursos*.

What has been transcribed as one sentence in English has been marked as two sentences in Portuguese. Since we did not annotate discourse relations between free adjuncts and their matrix clauses, the version in English has not been annotated with any discourse relations. But our Portuguese annotator quite rightly inserted an implicit conjunction *and*. We found that discrepancies of this type were ubiquitous across all languages, and conclude that they should be taken into account when conducting comparative statistical analyses, e.g. Kappa (Cohen, 1960), and Krippendorff's alpha (Krippendorff, 2004).

Turning now to translation, differences in how a translator interprets how the sentences and/or clauses of the source text relate to one another, can lead to differences in what sense relations annotators themselves can infer from a text, resulting in variations in sense labelling. For example consider the following token in English against its German translation:[1]

Talk 1971 (EN) As [***I watched people who I knew, loved ones, recover from this devastation***] arg1, [***one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses***] arg2. Sense = temporal:synchronous

Talk 1971 (DE) [***Ich beobachtete Menschen , die ich kannte , liebte , wie sie sich von dieser Verwüstung erholten***] arg1 , aber [***eine Sache quälte mich zutiefst , und zwar , dass viele der Amputierten in diesem Land nicht ihre Prothesen benutzten***] arg2 . Sense = concession:arg2 as denier

Here, the German translator has interpreted a concessional relationship between the arguments, and accordingly has used the connective *aber* (similar to the English *but*) to mark the relation. So even though we know the relation was originally intended as temporal, our German annotator was forced to mark it as concessional.

Moving onto sense relations – spoken communication with a co-present audience invites additional functions that utterances can serve. One common rhetorical device we observed was the speaker asking a question and then immediately answering it. To capture this sequence we created an additional sense relation – Q\A (marked as **type** = *implicit*, and **NONE** in the connective box). For example:

Talk 1927 (EN) [***Are investors, particularly institutional investors, engaged***] arg1? Well, [***some are, and a few are really at the vanguard***] arg2. Sense: Q\A

In the above token we also agreed that *well* has a pragmatic role, not a semantic one, i.e. *well* lets the hearer know that the question is being considered, and an answer will follow. Consequently we excluded *well* from the argument span, given that pragmatic inference is beyond the scope of our task. NB: *now* was also excluded from several tokens on the same principle.

[1] Thanks to Yulia for finding this example.

Lastly, given that PDTB is chiefly concerned with relations between adjacent sentences, we felt that many broader relations in the text were missed, i.e. relations with large gaps in between. For example consider the following:

> Talk 1927 (EN) *So how are companies actually leveraging ESG to drive hard business results? One example is near and dear to our hearts. In 2012, State Street migrated 54 applications to the cloud environment, and we retired another 85. We virtualized our operating system environments, and we completed numerous automation projects. Now these initiatives create a more mobile workplace, and they reduce our real estate footprint, and they yield savings of 23 million dollars in operating costs annually, and avoid the emissions o a 100,000 metric tons of carbon. That's the equivalent of taking 21,000 cars off the road.* [***So awesome, right***] arg1? [***Another example is Pentair***] arg2. Sense: NoRel

In this stretch of text it would make sense to put an implicit conjunction *and* between the first sentence "*One example is near and dear to our hearts."* and *"Another example is Pentair.",* but following rules on adjacency we simple marked no relation (NoRel) between the last two sentences. Marking broader relations could be a fruitful avenue of further research.

All in all, this STSM was a huge success; we carried out five adjudication sessions as described above, we updated and developed our set of guidelines specific to annotating TED Talks (or indeed any other form of prepared speech), we reached a high level of alignment across all languages, and we made a good start on developing a richly annotated corpus of parallel translated texts. It is our hope that the work will continue, and the corpora be extended to include additional languages.

Acknowledgments

Many thanks to Amalia Mendes for being a wonderful host, and for making me feel very welcome. Thanks also to Deniz Zeyrek for co-hosting, and showing me around the city (together with Amalia). Also many thanks to Bonnie Webber for encouraging me to apply for this STSM, helping to arrange it, providing feedback on the annotation task, and providing comments on this report. Thanks also to Yulia Grishina for providing material for this report, and thanks to all others involved with the project.

References

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Hovy, E., & Lavid, J. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, *22*(1), 13-36.

Krippendorff, K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411-433.