SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: IS1312
STSM title: Annotation of discourse relations in TED talks (38151)
STSM start and end date: 02/01/2018 - 16/01/2018
Grantee name: Giedre Valunaite Oleskeviciene

PURPOSE OF THE STSM

Lithuanian researchers are working on enriching the existing corpora and also are looking for ways to make the corpora inter-operable and co-searchable through the annotation of discourse-relational devices.

The aim of the STSM is extending the available resources and lexicons of discourse-relational devices in Lithuanian by cooperating with the international team of the researchers. The creators of TED-MDB (TED-Muliti-lingual Discourse Bank) (Zeyrek et al., 2018), Prof. Dr. Deniz Zeyrek and Dr. Amalia Mendes, who kindly invited more teams from TextLink to join their initiative of creating an open resource corpus TED-MDB and our Lithuanian team is eager to join. The aim is achieved by adding Lithuanian annotated texts to the existing TED-MBD corpus which already includes 6 languages: Turkish, English, Polish, German, Russian and Portuguese. The main focus of the STSM is working on the annotation of TED talks in Lithuanian. Joining the team of researches of Middle East Technical University and working on annotating TED talks is a great opportunity for me to grow as a researcher and expand the set of available resources in Lithuanian language looking for the ways making the resources co-searchable.

The aim of the STSM is closely related to one of the main objectives of TextLink action which is extending the set of available resources and lexicons of discourse-relational devices (DRDs).

During the STSM there are great opportunities for me to deepen my knowledge on PDTB annotation scheme, get acquainted with the most recent version of PDTB annotation scheme and learn from my colleagues the subtleties of annotating spoken-like texts of TED talks.

Detailed steps of the research include the following: first, getting acquainted with the newest version of PDTB annotation scheme, then, collecting the parallel TED talk texts in Lithuanian for annotation, finally, annotating the chosen texts and then comparing cross-linguistically with the view of looking for regularities and striking features.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

The development of the Penn Discourse Treebank (PDTB) (Prasad et al., 2016), which is a large-scale annotated corpus of discourse relations, fuelled interest in cross linguistic studies of discourse relations of similar PDTB-based annotation projects in other languages. Lithuanian texts were chosen and annotated with discourse relations (e.g., causal, contrastive, elaboration and temporal relations) by closely collaborating with Prof. Dr. Deniz Zeyrek, who was my host in this visit. Prof. Dr. Deniz Zeyrek is an expert on PDTB-based annotation, and provided very helpful insight and close collaboration. During the scientific mission it was possible to move towards the direction of developing the available resources of discourse-relational devices in Lithuanian, which would not be possible without the visit.

In the process of discourse relation annotation we observed the main principles of the PDTB framework, that the annotation approach should be theory-neutral and lexically grounded. Theory-neutral approach means that the annotation is not based on a specific discourse structure theoretical grounds. And lexically grounded perception implies that annotator judgements are effectively elicited. The annotation included both explicit and implicit discourse relations with their arguments, called Arg1 and Arg2, and the senses of discourse relations. Explicit discourse connectives include expressions from four grammatical classes: subordinating conjunctions, coordinating conjunctions, sentential relatives and discourse adverbials. The main task is to identify if the words and phrases function as discourse connectives as they can have other non-discourse functions. In the cases of annotating implicit connectives the annotator has to insert a connective that best expresses the inferred relation. The cases are annotated as entity relations when the only relational inference could be made that the second sentence identifies one or more entities from the previous sentence and describes the entity/entities. The Argument annotation follows the rule that the label Arg 2 is attached to the argument which appears in the clause that is syntactically bound to the connective and then the other argument is marked Arg 1. In the cases of hypophora related to question-answer annotation, we followed the natural order of the arguments marking questions as Arg1and the following answers Arg2. Besides, we did not have to use Altlex (alternative lexicalization) as there is a specific conjunction in the Lithuanian language *ar* which is used to formulate a questions. For example:

[Do] (AltLex) ***companies that take sustainability into account really do well financially****? The answer that may surprise you is yes.* (annotated Hypophora)

[Ar] įmonės, atsižvelgiančios į tvarumą, išties finansiškai sėkmingos? ***Jus galintis nustebinti atsakymas yra „taip“.*** (annotated Hypophora plus explicit Expansion: Level_of_detail: Arg2_as_detail)

I also had a chance to attend a talk by Dr. Umut Özge, a part of METU Cognitive Science Colloquium series. The speaker analysed the issues in the semantics of nominalization, primarily aiming in his talk to raise interest in the semantics of various nominalization processes through an informal discussion of some data from Turkish, where nominalization interacts with argument structure, existence presuppositions, temporal reference, and event structure. The talk raised my awareness on the processes of nominalization and preprograming linguistic cognition of AI.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

During the visit after annotating the chosen texts we explored the texts cross-linguistically with the view of looking for regularities and striking features. In PDTB annotation connectives are identified and then linked to their two arguments Arg 1 and Arg2, where usually Arg2 is syntactically integrated with the connective. What is more, links between adjacent sentences when no connective is present are also annotated in PDTB by identifying such cases as implicit connectives. All the relations both explicit (signalled by a connective) and implicit are given a sense label taken from the hierarchy of senses. While comparing the annotated texts cross-linguistically the main striking feature is observed that some implicit relations in English texts are translated explicitly in Lithuanian texts. For example:

that's okay, right [but] ***We want more*** (annotated **implicit** Comparison: Concession: Arg2_as_denier)

Nebogai, tiesa [Bet] ***mes norim daugiau*****.** (annotated **explicit** Comparison: Concession: Arg2_as_denier)

The example demonstrates that the implicit relation in the original text is reveled explicitly by Lithuanian contrastive conjunction bet. Such cases demonstrate, that translators are successfully able to render the existing implicit relations between the sentences explicitly.

However, there are cases when the explicit connectives are rendered implicitly which might lead to the loss of the sense annotated in the original text. For example:

only looking at race doesn't really contribute to our development of diversity. [So] ***if we're trying to use diversity as a way to tackle some of our more intractable problems, we need to start to think about diversity in a new way****.* (annotated **explicit** Contingency: Cause: Result)

žiūrėti tik į rasę nepadeda bandant prisidėti prie įvairumo vystymo. [taigi] ***Bandome įvairumą naudoti sprendžiant kai kurias sudėtingesnes problemas, turime pradėti kitaip galvoti apie įvairumą****.* (annotated **implicit** Contingency: Cause: Result)

[if] ***we're trying to use diversity as a way to tackle some of our more intractable problems****, we need to start to think about diversity in a new way.* (annotated **explicit** Contingency: Condition: Arg2_as_condition)

Bandome įvairumą naudoti sprendžiant kai kurias sudėtingesnes problemas, [todėl] ***turime pradėti kitaip galvoti apie įvairumą***. (annotated **implicit** Contingency: Cause: Result)

As it could be seen the translator chose not to render any of the two explicit connectives and the loss of the annotated senses in the original could be observed in Lithuanian text. If the sense of the result could be felt implicitly, however the sense of condition is totally lost. Finally, it could be concluded, that by comparing the annotated texts, the translators could gain certain insights on rendering the senses and observing the quality of translation.

FUTURE COLLABORATIONS (if applicable)

This scientific mission is a part of an ongoing collaboration within the framework of COST action IS1312 TextLink with my host in the Turkish University METU in Ankara. Prof. Dr. Deniz Zeyrek has an active role in the collaboration while creating of TED-MDB (TED-Muliti-lingual Discourse Bank). We explored the theoretic aspects and the guidelines of annotation of Lithuanian texts. This mission provided the means to closely collaborate on a day-to-day basis and extend my understanding of PDTB tool and exploit new perspectives of the research. The implemented short-term scientific mission reinforced the existing links between the Middle East Technical University and Mykolas Romeris University in Vilnius. It extend the initial work done through collaboration between the universities on extending the available resources and lexicons of discourse-relational devices and provided the perspectives for agreement on joint plans. During the visit I had the chance to meet other members of the staff from the Middle East Technical University. Particularly helpful were the valuable insights from Murathan Kurfali. The annotation and exploration of Lithuanian texts would not have been possible without the grant for the scientific mission. Our preliminary results are promising, and show that discourse annotation can help in dealing with translation of implicit relations and enrich possibly incomplete and inconsistent knowledge. In the future, we want to extend the results obtained during the visit and investigate the ways translators deal with implicit relations. During the limited period of the visit we only started working towards the identified direction. We would also like to evaluate the performance of translators by annotating and researching a broader scope of the texts.

References

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A Discourse-Annotated Corpus of Conjoined VPs. *LAW X*, 22.

Zeyrek, D., Mendes, A. & Kurfali (2018). M. Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank. Accepted for inclusion in LREC 2018, Miyazaki, Japan.