February 21, 2018

STSM Report
January 20 – January 30, 2018, Madrid
Deniz Zeyrek

This report summarizes the STSM visit that I undertook to collaborate with Dr. Julia Lavid Lopez at Universidad Complutense de Madrid on the extension of TED-Multilingual Discourse Bank (TED-MDB) to Spanish.

TED-MDB is an ongoing effort to develop a multilingual discourse corpus of TED talks. The corpus is a collection of TED talks transcripts (Cettalo et al., 2012) annotated at the discourse level in the PDTB style (Prasad et al., 2014). One of the aims of TED-MDB is to allow cross-linguistic comparisons; in particular, to reveal similarities or differences in how meanings are mapped onto discourse relations across languages, e.g. by explicit or implicit means. A second aim is to offer a multilingual corpus to the community where discourse connectives and other issues involved around discourse can be used as an input to NLP applications.

The TED-MDB corpus currently involves TED talks transcripts from six languages (Turkish, European Portuguese, Russian, German, Polish) and we believe it would benefit from the addition of transcripts from new languages, such as Spanish. During this visit,

- We discussed in detail how PDTB 3.0 annotation guidelines have been adopted to TED-MDB,
- We considered the translation quality of the Spanish transcripts,
- We analysed one of the Spanish TED talk transcripts in our database and Dr. Julia Lavid Lopez created discourse-level annotations on this text with the PDTB annotation tool.

In the rest of this report, I detail these points.

The data and the practice annotation procedure

One of the first things that need to be considered in the development of a multilingual corpus is the translation quality of the texts included for analysis and annotation. During the creation of TED-MDB, the transcripts in Turkish and European Portuguese were selected by going over each transcript making sure the texts had a good level of translation quality and were easy to comprehend (Zeyrek et al., 2018). Similarly, the first step of extending the TED-MDB to Spanish has been an analysis of the translation quality of the transcripts. In case the translation quality is not at the expected level, the teams involved in contributing to the corpus could consider retranslating the texts or choosing a set of other transcripts to translate professionally. A set of Spanish transcripts had been made available to us by Murathan Kurfalı from the WIT3 website (https://wit3.fbk.eu). Going through these texts we discussed their translation quality and we settled on one of the texts to practice with.

After discussing the PDTB and TED-MDB guidelines, Dr. Lavid started to annotate the text in the way described in Zeyrek (et al., 2018). In other words, she went through the text sentence by sentence and annotated both intra- and inter-sentential discourse relations that hold in the text. She used the PDTB annotation tool (Lee, et al. 2016), as in TED-MDB. Since the PDTB annotation tool was already introduced in the Workshop "Annotation of Discourse Relational Devices (DRDs): Multilingual and Multimodal Challenges" held in Universidad Complutense de Madrid in July 2017 (http://textlink.ii.metu.edu.tr/annotation-discourse-relational-devices-drds-multilingual-and-multimodal-challenges), we were able to launch the tool and quickly and easily. Dr. Julia Lavid Lopez used version 4.0 of the tool.

While Dr. Lavid Lopez created the annotations, I was with her to take her through the annotation procedure adopted in creating TED-MDB, as well as helping with technical issues that could arise in launching the tool or during the annotations. After she annotated each discourse relation, we discussed her choice of the discourse relation type and the sense in detail. Then, she recorded the annotation token and continued with the next relation in the text, and the cycle continued.

The sample annotations were created on the basis of the revised PDTB 3.0 sense hierarchy (Webber et al., 2016). The revised sense hierarchy maintains the four top-level senses of PDTB 2.0 (Prasad et al., 2007), namely Expansion, Temporal, Contingency and Comparison. The updates mainly involve the addition of new subsenses that were missing in the previous version (e.g. Contingency:purpose) and an attempt to simplify the sublevels, as in the reduction of certain sublevels of Expansion. Given PDTB's lexically based approach to discourse, each relation type (whether it is explicit, implicit, or alternative lexicalization) needs to be assigned a sense (or multiple senses) from the sense hierarchy. Dr. Lavid Lopez annotated all major types of discourse relations (explicit relations, implicit relations, entity relations, etc.) together with their binary arguments and senses, where relevant. We concentrated on annotating the relations with a single sense, though in some cases we discussed the possibility of a secondary sense that could also be added.

Observations

In general, explicit relations (example (1)), alternative lexicalizations (example (2)) have been quite easy to notice and annotate in the Spanish text we looked at.

(1) *Nací y crecí en Sierra Leona, un país pequeño y muy hermoso de África occidental. Un país rico tanto en recursos materiales como en talento creativo*. Sin embargo, **Sierra Leona es tristemente conocida por la guerra insurgente de 10 años, de los 90, en la que poblados enteros fueron reducidos a cenizas**. (Comparison: Concession)

I was born and raised in Sierra Leone, a small and very beautiful country in West Africa, a country rich both in physical resources and creative talent. However, **Sierra Leone is infamous for a decade-long rebel war in the '90s when entire villages were burnt down**. (Comparison: Concession: Arg2_as_denier)

(2) *Al ver a la gente que conocía, gente querida, recuperarse de esta devastación, algo que me preocupaba profundamente era que muchos de los amputados del país no iban a usar sus prótesis*. El motivo, **luego me daría cuenta, era que los encajes ortopédicos producían dolor porque no calzaban bien.** (Contingency:Cause:Reason)

As I watched people who I knew, loved ones, recover from this devastation, one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses. The reason, **I would come to find out, was that their prosthetic sockets were painful because they did not fit well.** (Contingency:Cause:Reason)

Entity relations (example (3)) have been annotated in the Spanish text, though we believed these relations were not clearly specified in guidelines and could not be consistently annotated even by an expert annotator.

(3) *El encaje ortopédico es la parte donde el amputado introduce lo que quedó de su miembro uniéndolo al tobillo ortopédico*. **Incluso en los países desarrollados, a un paciente le lleva de 3 semanas a varios años conseguir un encaje cómodo y a veces no lo consigue nunca**. (EntRel)

The prosthetic socket is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle. **Even in the developed world, it takes a period of three weeks to often years for a patient to get a comfortable socket, if ever**. (EntRel)

Implicit relations are differently realized in English and Spanish but that is to be expected and interesting from a contrastive perspective. For example, while example (4) is annotated as an implicit concession relation in the Spanish text, it is annotated as an alternative lexicalization in the English version.

(4) *Muchas veces estos encajes generan niveles de presión insoportables en los miembros del paciente, provocándoles úlceras y ampollas*. (IMPLICIT=pero) **No importa lo buena que sea la prótesis**. (Comparison: Concession)

Such sockets often leave intolerable amounts of pressure on the limbs of the patient, leaving them with pressure sores and blisters. It does not matter **how powerful your prosthetic ankle is**. (Comparison: Concession:Arg2-as-denier)

Yet another common situation we find in the annotated texts is missing relations in either the Spanish version or the English text; for example in (5), an implicit discourse relation is annotated in the Spanish text, whereas no discourse relation has been annotated for the English version. But an explicit intra-sentential relation is annotated in both texts (see examples (6) and (7)).

(5) *En este período, se calcula que 8000 hombres, mujeres y niños sufrieron amputaciones de sus brazos y piernas*. (IMPLICIT=por ello) **Mientras escapábamos con mi familia de uno de esos ataques, a los 12 años decidí hacer todo lo posible para que mis hijos no tuvieran que pasar por situaciones como esa**. (Contingency:Cause:Result)

An estimated 8,000 men, women and children had their arms and legs amputated during this time. As my family and I ran for safety when I was about 12 from one of those attacks, I resolved that I would do everything I could to ensure that my own children would not go through the same experiences we had. (No relation annotated.)

(6) As **my family and I ran for safety when I was about 12 from one of those attacks**, *I resolved that I would do everything I could to ensure that my own children would not go through the same experiences we had*. (Temporal:Synchronous)

(7) Al **ver a la gente que conocía, gente querida, recuperarse de esta devastación**, *algo que me preocupaba profundamente era que muchos de los amputados del país no iban a usar sus prótesis.* (Temporal:Synchronous)

Examples (8) and (9) illustrate the cases where no discourse relation has been annotated in the Spanish text due to the different translation of the sentence:

(8) El encaje ortopédico es la parte donde el amputado introduce lo que quedó de su miembro uniéndolo al tobillo ortopédico.

The prosthetic socket is the part in which the amputee inserts their residual limb, and **which connects to the prosthetic ankle**. (Expansion:Conjunction)

(9) Entonces un día, cuando conocí al profesor Hugh Herr hace unos dos años y medio, me preguntó si sabía cómo resolver este problema. Le dije: "Todavía no, pero me encantaría descubrirlo."

So *one day, when I met professor Hugh Herr about two and a half years ago*, and **he asked me if I knew how to solve this problem**, I said, "No, not yet, but I would love to figure it out." (Expansion:Conjunction)

Finally, we considered the interactive nature of TED talks and the extension of PDTB guidelines to capture this in the TED-MDB corpus. TED talks are prepared, scripted, very well-rehearsed and eventually delivered as a monologue to a live audience. They usually revolve around a dramatic story and tend to spread provocative ideas (http://speakupforsuccess.com/public-speaking-tip-82-dont-be-intimidated-by-the-ted-talk-style/). They represent a genre involving various discourse modes. While preserving the formal aspect of the speech event, speakers integrate the interactive aspects of spoken discourse to their speech. Thus, annotating the interactive nature of TED talks is a challenge for the TED-MDB corpus. For example, speakers often integrate personal histories, as in example (9) above. While TED-MDB annotates discourse relations within such narratives, the relation of the personal narrative within the larger discourse is not shown, as it lies out of the scope of our annotation framework. On the other hand, the relationship of attribution expressions such as "I said" in example (9) and the attributed text span itself ("no not yet but…") is left for further research.

Conclusion and looking ahead

To conclude, working with Dr. Julia Lavid Lopez on a Spanish transcript to create sample discourse-level annotations has revealed most of the issues an annotator would face in dealing with a multilingual corpus. On the annotation side, the major discourse relation types TED-MDB aims to capture are often easy to capture in Spanish too, although clear guidelines seemed necessary to distinguish implicit relations from alternative lexicalizations or entity relations. On the theoretical side, we were able to consider a range of issues related to discourse, e.g. the way intra-sentential connectives existing in the English texts are translated into Spanish, the nature of entity relations and how to recognize them, and the role of attribution in TED talks. In the near future, we plan to submit a paper to a relevant NLP conference and share our observations with the community.

References

Cettolo, M., et al. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. In *Proc. of EAMT*, pp. 261-268, Trento, Italy.

Lee, A., Prasad, R., Webber, B., & Joshi, A. K. (2016). Annotating Discourse Relations with The PDTB Annotator. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations (pp. 121-125).

Prasad, et al. (2014) Reflections on the Penn Discourse Treebank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. L. (2007). The penn discourse treebank 2.0 annotation manual.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 22-31).

Zeyrek, D., A. Mendes, S. Gibbon, Y. Grishina, M. Ogrodniczuk & Kurfali, M. (in preparation) TED-Multilingual Discourse Bank (TED-MDB): TED Talks annotated in the PDTB style.

Zeyrek, D. & Mendes, A. (2017) Multilingual extension of PDTB annotation: the case of TED Multilingual Discourse Bank. *Textlink, Multilingual PDTB Annotation Workshop*, (June 8-12, 2017), Universidade Complutense, Madrid.

Zeyrek, D. (2017). TED Multilingual Discourse Bank (TED-MDB): A parallel annotated in the PDTB style (Invited talk). 11th Linguistic Annotation Workshop (LAW) , European Chapter of the Association of Computational Linguistics. April 3rd, 2017, Valencia.

Zeyrek, D., Mendes, A., Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank. In: LREC.