# An automatic discourse relation alignment experiment on TED-MDB

**Sibel Özer**
Graduate School of
Informatics, Middle East
Technical University,
Ankara, Turkey
e159606@metu.edu.tr

**Deniz Zeyrek**
Graduate School of
Informatics, Middle East
Technical University,
Ankara, Turkey
dezeyrek@metu.edu.tr

## Abstract

This paper describes an automatic discourse relation alignment experiment as an empirical justification of the planned annotation projection approach to enlarge the 3600-word multilingual corpus of TED Multilingual Discourse Bank (TED-MDB). The experiment is carried out on a single language pair (English-Turkish) included in TED-MDB. The paper first describes the creation of a large corpus of English-Turkish bi-sentences, then it presents a sense-based experiment that automatically aligns the relations in the English sentences of TED-MDB with the Turkish sentences. The results are very close to the results obtained from an earlier semi-automatic post-annotation alignment experiment validated by human annotators and are encouraging for future annotation projection tasks.

## 1 Introduction

There has been a recent interest in creating discourse-level annotations in multilingual corpora. Some efforts include Popescu-Belis et al. (2012), Stede et al. (2016), Samy et al. (2008), and Zufferey et al. (2017), annotating multilingual corpora for discourse-level phenomena such as coreference and discourse relations (DRs). Given the need for quickly building multilingual corpora where human effort is reduced, this paper aims to describe the initial steps to enlarge TED-Multilingual Discourse Bank (TED-MDB) (Zeyrek et al., 2019), ultimately using an annotation projection approach.

The notion of DR refers to semantic linkages that hold between text spans with labels as comparison, contingency, elaboration. DRs may exist both within and across sentences (examples 1 and 2):

(1) Since my neighbours are away, their lights are off.

(2) The high school student writes wonderful essays. In addition, she is very good in math.

In (1) and (2), *since* and *in addition* make the discourse relations salient; the relations instantiated by these expressions are called explicit relations. In many cases, however, a relation may lack an overt explicit connective. These have been known as implicit relations.

TED-MDB annotates TED talks for discourse relations in the original language, English as well as the transcripts of 5 languages (German, Polish, Portuguese, Turkish and Russian) by following the principles of the PDTB. Thus, five DR types (Explicit, Implicit, AltLex, EntRel, NoRel) are annotated together with their binary arguments and senses, where appropriate. For sense assignment, the PDTB-3 sense hierarchy is used (Webber, et al., 2016). TED-MDB ultimately aims to provide a clearly described level of discourse structure and semantics in multiple languages and engender discourse parsing studies in multiple languages. But the resource is still small. In order to reduce human effort and to quickly enlarge this corpus, annotation projection appears to be a viable option.

Discourse relation annotation projection has been a recent practice. For example, using a parallel English-German corpus, Versley (2010) attempted to disambiguate German connectives via projection. Laali et al. (2017) projected DR annotations from English texts onto French texts on English-French parallel texts from Europarl.

TED-MDB annotations have been created by native speaker annotators independently of the annotators of other languages. Despite this procedure, a post-annotation semi-automatic alignment experiment validated by humans showed a good alignment performance. To move towards annotation projection and provide an empirical justification for this endeavour, the present study has two distinct but related aims: (a) to create a large parallel corpus of English-Turkish TED talk bi-sentences that would serve as the basis for extending TED-MDB, and (b) as a proof-of-concept experiment, to automatically align all 6 English texts of TED-MDB with the equivalent Turkish transcripts to understand the extent to which a fully automatic alignment reaches the results of the previous human-validated experiment. We hypothesize that the closer the results are to each other, the higher the chance of a successful DR projection would be.

## 2 A Parallel corpus of English-Turkish TED talk transcripts

To create the parallel corpus, first, English TED talk transcripts and their Turkish versions uploaded to the TED Talks website[1] until 25th of March, 2019 were downloaded using a web crawler implemented in Python. This resulted in a total of 2977 texts in English and their translated versions in Turkish. In the pre-processing stage, along with other steps taken such as the correction of the wrong HTML codes, eliminating unnecessary spaces and apostrophe marks, texts which contain song lyrics or texts which present musical performances were removed since such texts also included musical notes and created noise. One English file which has missing transcripts in Turkish was also discarded. The corpus has a total of 2852 files in each language in the end (Table 1).

| | Doc. Count | Paragraph Count | Sentence Count | Word Count | Token Count |
|---|---|---|---|---|---|
| English | 2852 | 5704 | 341.574 | 5.560.816 | 6.411.236 |
| Turkish | 2852 | 5704 | 348.617 | 3.937.529 | 4.682.604 |

Table 1: Data statistics of English-Turkish parallel TED talks

In the second step, each text file was tokenized into words and punctuation marks and English-Turkish bi-sentences were drawn. The tokenizer and the sentence aligner in the Uplug tool[2] were used (Tiedemann, 2003). Manual corrections were done on misaligned bi-sentences, which were mostly due to missing punctuation marks on either language or different translations in Turkish, amounting to misalignments in almost two thirds of the documents in the Turkish part. Such errors were corrected manually, a process that took approximately 4 weeks of intense work. In this way, a parallel corpus containing 325.398 bi-sentence units were obtained.

## 3 Proof-of-experiment and evaluation

To align the documents at the DR level, we first transferred the annotations onto the base text files of both languages and performed word- and punctuation-tokenization as well as sentence alignment on these documents and did manual corrections.[3]

The first step of the experiment involves assigning a sense/type score to the bi-sentences. We used a ranking algorithm that depends on the DR type, sense and argument spans – this we call the sense/type score. To score the DRs in each bi-sentence, we first pair them with respect to all 5 DR types and all three sense levels (class, type and subtype levels) of the PDTB-3 sense hierarchy. We assign a sense/type score of 1 or 0 (match or mismatch) to each bi-sentence on four criteria (Table 2). At the second phase, we add Bleu scores to the sense/type scores through the following steps: if Level1 senses of the DR pairs match, we translate the English arguments into Turkish and calculate the Bleu score for each translated argument and the original Turkish argument (arg1EnT[4] -arg1Tr, arg1EnT-arg2Tr, arg2EnT - arg1Tr, arg2EnT-arg2Tr).[5] The process is repeated by translating the Turkish arguments into English

[1] https://www.ted.com

[2] https://github.com/Helsinki-NLP/Uplug

[3] This was necessary as sentence-alignment was not performed on the TED-MDB texts before.

[4] T stands for translated version

[5] The translation phase was necessary to ensure the correct alignment of the argument spans as well as their content.

and assigning them Bleu scores. Maximum Bleu scores of each process are selected, summed and added to the sense/type scores. All translations are done using Google Translate API.

In example 3, three DRs are instantiated by the discourse connectives *ama* ‘but’,*gibi* ‘as’, *ve* ‘and’. The connectives and their sense tags on the English side are paired with their Turkish counterparts constituting 9 doublets as shown in (4).

(3) Years have passed, but many of the adventures I fantasized about as a child -- traveling and weaving my way between worlds other than my own — have become realities through my work as a documentary photographer. But no other experience has felt as true to my childhood dreams as living amongst and documenting the lives of fellow wanderers across the United States. (TED Talk no. 2009)

Yıllar geçti, ama çocuk olarak hayalini kurduğum birçok macera -- benim dünyam dışındaki dünyalar arasında seyahat ederken ve yoluma dokunurken -- bir belgesel fotorafçısı olarak işim aracıyla bunlar gerçek oldu. Ama hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak kadar ve Birleşik Devlet boyunca gezgin arkadaşların arasında yasamak kadar gerçek hissettirmedi.

(4) *English*: ( DR _ Explicit _ S1 _ Comparison.Concession.Arg2-as-denier _ DC _ **But** ) - ( DR _ Explicit _ S1 _ Comparison.Similarity _ DC _ **as** ) - ( DR _ Explicit _ S1 _ Expansion.Conjunction _ DC _ **and** )
*Turkish*: ( DR _ Explicit _ S1 _ Comparison.Concession.Arg2-as-denier _ DC _ **Ama** ) - ( DR _ Explicit _ S1 _ Comparison.Similarity _ DC _ **kadar** ) - ( DR _ Explicit _ S1 _ Expansion.Conjunction _ DC _ **ve** )

We assign each member of the doublet a sense/type score and a Bleu score. In Table 2, while *Ama* matches *But* in all four criteria and receives 1s, it matches as in two criteria*:* Level1 sense and the DR type. The DR pair that has the maximum score is selected as an aligned pair. In this case, *But-Ama, as-kadar,* and *and-v*e pairs are selected.

| | Ama | kadar | ve |
|---|---|---|---|
| But | 1111+90 | 1001+60 | 0 |
| as | 1001+50 | 1101+80 | 0 |
| and | 0 | 0 | 1101+50 |

Table 2: Scoring Table (the first score shows the sense/type score, the second score is the Bleu score)

Finally, we calculate precision, recall and F1 score by taking the English annotations as gold. We obtained an average F1 score of 0.80 distributed over the bi-sentences of 6 documents. This is lower than the F1 score of 0.84 in Zeyrek et al. (2019), but can still be considered promising.

Table 3 shows that the majority of explicit connectives in TED-MDB are rendered as explicit or Altlex relations in Turkish. This suggests that a DR annotation projection task involving the explicits and AltLexs could give better results than implicits or other DR types.

| | | | | | **Turkish** | | | |
|---|---|---|---|---|---|---|---|---|
| | | **AltLex** | **EntRel** | **NoRel** | **Explicit** | **Implicit** | **NA** | **Eng.To-** |
| **English** | **AltLex** | 21 | 0 | 0 | 8 | 6 | 11 | 46 |
| | **EntRel** | 0 | 61 | 0 | 0 | 0 | 17 | 78 |
| | **NoRel** | 0 | 0 | 42 | 0 | 0 | 7 | 49 |
| | **Explicit** | 15 | 0 | 0 | 192 | 25 | 57 | 289 |
| | **Implicit** | 5 | 0 | 0 | 15 | 131 | 43 | 194 |
| | **NA** | 11 | 9 | 9 | 61 | 40 | | |
| | **Tur.To-** | 59 | 70 | 51 | 276 | 202 | | |

Table 3: Distribution of Discourse Relations in TED-MDB

## 4 Conclusion

In this paper, we first constructed a large parallel corpus of English-Turkish TED talk transcripts. Secondly, we performed a sense-based automatic DR alignment experiment where we aligned the DRs created for the English part of TED-MDB with the Turkish part. We obtained an F1 score close to the F1 score of an earlier semi-automatic post-alignment experiment validated by humans. Based on this score and the fact that English explicits and AltLexes are mostly rendered similarly in Turkish, we plan to automatically annotate new English transcripts using a shallow discourse parser and project explicit and AltLex annotations onto the Turkish part. In the future, the approach will be tested for more language pairs.

## References

Bentivogli, L., & Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. Natural Language Engineering, 11(3), 247-261.

Laali, M., & Kosseim, L. (2017). Improving Discourse Relation Projection to Build Discourse Annotated Corpora. arXiv preprint arXiv:1707.06357.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B., & Zufferey, S. (2012). Discourse-level annotation over europarl for machine translation: Connectives and pronouns(No. CONF).

Samy, D., & González-Ledesma, A. (2008). Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic-Spanish-English). In LREC.

Stede, M., Afantenos, S. D., Peldszus, A., Asher, N., & Perret, J. (2016). Parallel Discourse Annotations on a Corpus of Short Texts. In LREC.

Tiedemann, J. (2003). Recycling translations: Extraction of lexical data from parallel corpora and their application in natural language processing (Doctoral dissertation, Acta Universitatis Upsaliensis).

Versley, Y. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In Workshop on the Annotation and Exploitation of Parallel Corpora (AEPC), NODALIDA, Tartu, Estonia.

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., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2019). TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. Language Resources and Evaluation, 1-27.

Zufferey, S., & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. Corpus Linguistics and Linguistic Theory, 13(2), 399-422.