Visualizing Cross-Lingual Discourse Relations in Multilingual TED Corpora

Zae Kim, Vassilina Nikoulina, Dongyeop Kang, Didier Schwab, Laurent Besacier

Visualizing Cross-Lingual Discourse Relations in Multilingual TED Corpora

Zae Myung Kim[1,3], Vassilina Nikoulina[1], Dongyeop Kang[2], Didier Schwab[3], Laurent Besacier[1]
[1]NAVER LABS Europe
[2]University of Minnesota
[3]Univ. Grenoble Alpes, CNRS, LIG
{zae-myung.kim,vassilina.nikoulina,laurent.besacier}@naverlabs.com
dongyeop@umn.edu
didier.schwab@univ-grenoble-alpes.fr

Abstract

This paper presents an interactive data dashboard that provides users with an overview of the preservation of discourse relations among 28 language pairs. We display a graph network depicting the cross-lingual discourse relations between a pair of languages for multilingual TED talks and provide a search function to look for sentences with specific keywords or relation types, facilitating ease of analysis on the cross-lingual discourse relations.

1 Introduction

Discourse relations in texts describe how two discourse segments–which can be phrases, sentences, or even paragraphs–are connected logically to each other. For example, given the following two sentences, *"I have cooked some pasta."* and *"I am hungry."*, we may infer that the latter is a `cause` for the former, in which case, the discourse relation between them can be classified as an `implicit causation`.[1] Now, when we translate these sentences into another language, the grammar and pragmatics of that language may prefer a certain discourse relation over the others that may not be the same as that of the original language. For example, in the case of Korean, which is an agglutinative language, the heavy usage of Korean postpositions (particles) dictates that the `causal` relation in the above example is revealed `explicitly` (regardless of the usage of explicit discourse connectives).

Such a cross-lingual discourse analysis can provide some insights for improving the quality of (machine) translation (Meyer and Webber, 2013; Meyer and Poláková, 2013; Guzmán et al., 2014; Iruskieta et al., 2015; Chen et al., 2020). This paper presents an interactive system that visualizes the cross-lingual discourse relations using two human-annotated multilingual discourse datasets (Zeyrek et al., 2019; Long et al., 2020) derived from TED talks following the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) framework. Our interactive data dashboard provides users with an overview of relation preservation among all language pairs. We also display a graph network depicting the cross-lingual discourse relations between a pair of languages for a given TED talk, and provide a search function to find sentences with specific query or relation types. We believe that, with this visualization tool, users can easily browse through the multilingual talks and discover which parts of the talks share similar discourse relations, and where they diverge.

2 Preprocessing Datasets

We abbreviate the two discourse-annotated datasets, Zeyrek et al. (2019) and Long et al. (2020), as `TED-Mult` and `TED-ZH`, respectively. `TED-Mult` involves seven languages (English, German, Lithuanian, Polish, Portuguese, Russian, and Turkish); and `TED-ZH`, two languages (English and Chinese). Thus, our analysis is conducted across the 28 possible language pairs.

In order to facilitate cross-lingual analysis, we first performed sentence level alignment using a sentence segmentation tool[2] based on conditional random fields and a set of heuristic rules. We note that the reason for using this particular tool is that it guarantees the reconstruction of original texts when segmented sentences are joined together; and that it supports custom regex-based rules to add to allowlist or denylist if need be.

Once the TED talk scripts for all eight languages were segmented into sentence level, we cross-lingually matched the English sentences with sentences from other languages by the following procedure:

1. Using a multilingual Sentence-BERT model

[1]It is *implicit* because these sentences are not connected with an explicit discourse marker such as *"Because"*.

[2]https://github.com/zaemyung/sentsplit

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 165–170
November 10–11, 2021.

	Sentences	Annotations		Relation Types					Level-1 Senses				
		Intra.	Inter.	Exp.	Imp.	AltLex	EntRel	NoRel	Temp.	Cont.	Comp.	Exp.	Other
Chinese	173	217	163	135	122	46	46	27	40	88	49	126	0
English	397	314	365	299	202	48	81	49	44	133	73	288	10
German	408	186	379	242	216	18	59	30	31	119	57	260	9
Lithuanian	403	385	370	379	246	18	79	33	46	167	82	339	9
Polish	430	198	380	217	200	5	104	52	43	108	82	183	6
Portuguese	386	271	358	272	256	30	38	33	54	139	71	287	7
Russian	408	179	386	237	221	20	57	30	30	111	56	269	12
Turkish	424	332	408	336	213	67	72	52	41	162	75	330	7

Table 1: Table of corpora statistics for `TED-Mult` and `TED-ZH`.

(Reimers and Gurevych, 2019, 2020), fixed-size sentence embeddings were computed for all sentences of the eight languages.

2. In addition, non-English sentences were machine-translated into English, and their English sentence embeddings were computed using a monolingual English Sentence-BERT model.
3. For each TED talk, cosine similarity scores were computed between English and non-English sentences using the multilingual sentences embeddings, and also the monolingual sentence embeddings, separately.
4. Similarly, chrF scores (Popović, 2015) were computed between the English and the machine-translated (English) sentences.
5. We note that all of these scores were calculated for a fixed-size window of sentences as the cross-lingual sentences should be within a close range to each other.
6. Each type of scores was computed in both direction, i.e., `English` to `Lang.X` and `Lang.X` to `English`, and only their intersection of matched groups was kept.
7. Finally, the three types of scores were filtered by empirically defined thresholds and majority voting was used to pick the final matching groups of cross-lingual sentences for `Lang.X:English` pair for each TED document.

When aligning two non-English documents (`Lang.X:Lang.Y`), we used the alignments with the English (`Lang.X:English` and `Lang.Y:English`) as a pivot to cross-lingually match them.

To ensure the quality of the resulting alignments, we checked and revised each of them manually. We note that the manual correction was seldom required.

In total, we used seven cross-lingual talks from `TED-Mult` and four talks from `TED-ZH`, which are aligned with other languages. Overall, there are about 400 sentences per language, and Table 1 shows more details on the corpora statistics.

Level-1	Level-2	Level-3
Temporal	Synchronous	-
	Asynchronous	Precedence
		Succession
Contingency	Cause	Reason
		Result
		Neg. Result
	Cause+Belief	Reason+Belief
		Result+Belief
	Cause+Speech Act	Reason+Speech Act
		Result+Speech Act
	Condition	Arg1-as-Cond
		Arg2-as-Cond
	Condition+Speech Act	-
	Negative-Condition	Arg1-as-Neg. Cond
		Arg2-as-Neg. Cond
	Negative-Condition+ Speech Act	-
	Purpose	Arg1-as-Goal
		Arg2-as-Goal
Comparison	Concession	Arg1-as-Denier
		Arg2-as-Denier
	Concession+Speech Act	Arg2-as-Denier+Speech Act
	Contrast	-
	Similarity	-
Expansion	Conjunction	-
	Disjunction	-
	Equivalence	-
	Exception	Arg1-as-Excpt
		Arg2-as-Excpt
	Instantiation	Arg1-as-Instance
		Arg2-as-Instance
	Level-of-Detail	Arg1-as-Detail
		Arg2-as-Detail
	Manner	Arg1-as-Manner
		Arg2-as-Manner
	Substitution	Arg1-as-Subst
		Arg2-as-Subst

Table 2: Table of PDTB-3's hierarchical senses. In this paper, we consider up to the level-2 senses.

Figure 1: A screenshot showing the *"Overall Patterns"* of discourse relations among language pairs.

3 System Description

Our system is implemented using *Streamlit*[3] which is an open-source web application framework in Python. The aligned dataset with codes[4] and the demo system[5] are available online.

The system is organized into three sections, *"Overall Patterns"*, *"Pairwise Talks"*, and *"Search"*, where each section can be accessed via a navigation panel on the left-hand side.

3.1 Overall Patterns

In PDTB-3 framework, a discourse relation is represented by four components: relation type, level-1, 2, and 3 senses. As shown in Table 2, each subsequent sense level contains more fine-grained classes for the previous level, and we consider up to level-2 senses in our system.

The overall pattern in the preservation of cross-lingual discourse relations is depicted in a heatmap shown in Figure 1. We compute how many relations are exactly matched across the language pairs, considering relation type and levels of senses. For each language pair, from left to right, each cell denotes the accuracy of matching (1) relation type (implicit or explicit); (2) level-1 senses; and (3) level-1 and 2 senses jointly. The accuracy of rela-

[3] https://streamlit.io/

[4] https://github.com/zaemyung/Visualizing-Cross-Lingual-Discourse-Relations

[5] https://share.streamlit.io/zaemyung/visualizing-cross-lingual-discourse-relations/main/main.py

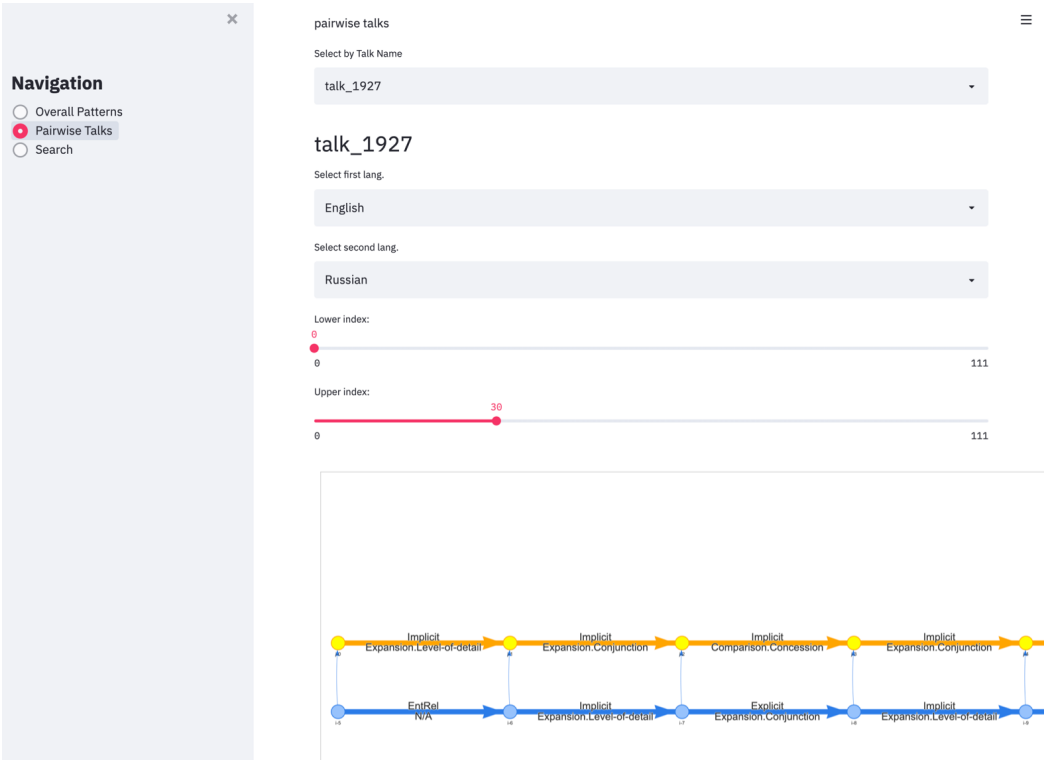

Figure 2: A screenshot showing the cross-lingual relation network graph of *"Pairwise Talks"*.

tions for cross-lingual alignments is computed as: $\frac{2*|I \cap J|}{|I|+|J|}$, where I and J are the set of corresponding relations for each alignment pair.

We observe that the relation type and level-1 senses are quite well matched across the language pairs, averaging 77% and 68%, respectively. The accuracy scores drop when we compare more fine-grained level-2 senses (48%).[6]

We note that, while the scores for Figure 1, were computed considering both the intra- and inter-sentential relations, it is possible to conduct the computation separately as well.

Below the heatmap, we have the top-scoring rules mined from association rule mining (Agrawal et al., 1993) where the algorithm is applied to a collection of language-specific occurrences of relations for a given language pair. While most of the rules learned are "identity" (same discourse relation preserved across language pairs), there are some interesting non-identical rules as well. For example, it is often observed that with high confidence (0.59), level-1 sense `Contingency` for Russian co-occurs with level-1 sense `Expansion` for English.[7]

At the end of the page, we present pie charts that illustrate the proportion of each relation type and level-1 and -2 senses of a source language being transferred to that of a target language, showing more fine-grained information than the heatmap. For example, in Figure 1, the left pie chart for the level-1 senses (`fist_cls`) depicts that the English's `Expansion` sense is mostly preserved (78.4%) in the corresponding Russian sentences. Similarly, the right pie chart shows that the English's `Implicit` relation type is kept the same in the Russian sentences 80.9% of the time across all talks.

These analyses on overall patterns can show how and where the discourse relations are preserved or diverged cross-lingually. From observations, we speculate about some of the causes for the divergence to be as follows: (1) The grammar and pragmatics of one language may favor the usage of one discourse relation over the other. (2) In some cases, especially when a discourse relation is `Implicit`, more than one relation could be possible. For

[6]We note that the avg. F1 score of matching Rhetorical Structure Theory (RST) relations across English, Spanish and Basque is 55% (Iruskieta et al., 2015).

[7]Both *confidence* and *lift* are two useful concepts in association rule mining that select interesting rules from the set of all possible rules. Given a rule, $A \rightarrow B$, the former (ranging from 0 to 1) computes the probability of seeing the consequent in a transaction given that it also contains the antecedent; the latter (ranging from 0 to ∞), measures how much more often the antecedent and consequent of the rule occur jointly than if they were statistically independent ($lift = 1.0$).

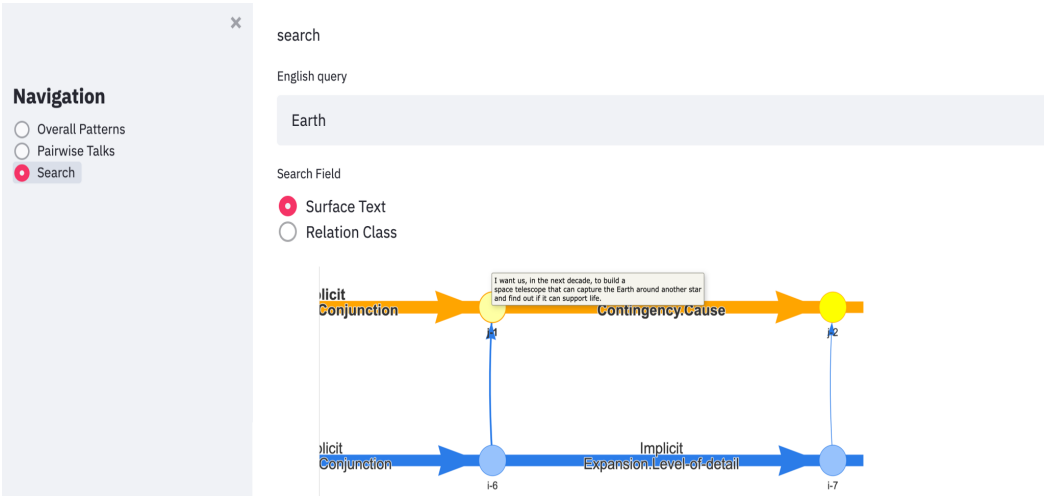

Figure 3: A screenshot showing the results of *"Search"* function where a sentence containing a specific English query is returned along its immediate neighboring sentences.

example, we often observed cases where the `Implicit.Expansion.Conjunction` and `Implicit.Comparison.Concession` relations were quite similar to each other.[8]

3.2 Pairwise Talks

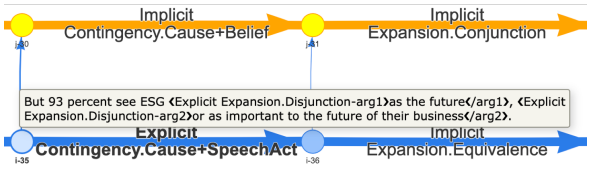

Figure 4: A screenshot showing an intra-sentential relation within a node.

For each TED talk and a pair of source and target languages, an interactive graph network is rendered to show cross-lingual discourse relations (Figure 2). A node in the graph represents a sentence, while an edge depicts inter-sentential discourse relation between two nodes. In the graph of the screenshot, the orange graph represents the discourse graph for the first five sentences in the Russian document, while the blue one shows that of the English counterpart. The vertical line between the two discourse graphs indicates cross-lingual sentence alignment. We note that the context window of the generated graph can be adjusted by the *Lower index* and *Upper index* sliders, in the case of really long TED talks where, otherwise, the resulting graph would be too large to render efficiently.

With this graph representation, we can easily view which nodes share similar cross-lingual discourse relations. For example, in the screenshot, we observe that the second relation between the second and third sentences is `Implicit.Expansion.Conjunction` for Russian and `Implicit.Expansion.Level-of-detail` for English. In this case, for both languages, the relation types and level-1 senses are `Implicit` and `Expansion`, respectively, while the level-2 senses are different (`Conjunction` versus `Level-of-detail`). In addition, by clicking on the node, we can observe any intra-sentential discourse relation present in the node by tagged spans in texts as illustrated in Figure 4.

3.3 Search

Users can look for a set of sentences that contain specific keywords or certain discourse relations on the *"Search"* page as displayed in Figure 3. Currently, the search is based on an English keyword, and it retrieves all the other multilingual sentences that are cross-lingually aligned with the English ones containing the keyword. The results include both the previous and the subsequent neighboring sentences to show the context around the query sentence.

4 Conclusion

We presented an interactive system that visualizes the cross-lingual discourse relations among multilingual TED talks, along with other related statistics. As future work, we plan to integrate an online multilingual discourse parser and a multilingual machine translation model into the system so that users can translate and analyze any given document and its translation on the fly.

[8]These kinds of issues were resolved via majority voting or discussion when creating the annotated datasets by their respective authors.

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber, and Deyi Xiong. 2020. Shallow discourse annotation for Chinese TED talks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1025–1032, Marseille, France. European Language Resources Association.

Thomas Meyer and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.