INTERNATIONAL CONFERENCE

RECENT ADVANCES IN

NATURAL LANGUAGE PROCESSING

D R A F T
P R O C E E D I N G S

Edited by
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Ivelina Nikolova, Irina Temnikova

Varn, Bulgaria

2–8 September, 2017

Preface

Welcome to the 11th International Conference on "Recent Advances in Natural Language Processing" (RANLP 2017) in Varna, Bulgaria, 2-8 September 2017. The main objective of the conference is to give researchers the opportunity to present new results in Natural Language Processing (NLP) based on modern theories and methodologies.

The conference is preceded by two days of tutorials (5–6 September 2015) and the lecturers are:

- Ondrej Bojar and Jindrich Helcl (Charles University)
- Noa Cruz (Virgen del Rocio Hospital, Seville)
- Veronique Hoste and Orphee De Clercq (Ghent University)
- Sanja Štajner (Mannheim University)

The conference keynote speakers are:

- Robert Dale (Language Technology Group)
- Josef van Genabith (DFKI, Saarbruecken)
- Veronique Hoste (Ghent University)
- Roberto Navigli (Sapienza University of Rome)
- Joakim Nivre (Uppsala University)

This year 25 regular papers, 40 short papers, 40 posters, and 2 demos have been accepted for presentation at the conference. In 2017 RANLP hosts five workshops on influential NLP topics: Natural Language Processing and Information Retrieval; Knowledge Resources for the Socio-Economic Sciences and Humanities; Human-Informed Translation and Interpreting Technology; Biomedical Natural Language Processing; and Language Technology for Digital Humanities in Central and (South-)Eastern Europe.

The proceedings cover a wide variety of NLP topics, including but not limited to: deep learning; machine translation; opinion mining and sentiment analysis; semantics and discourse; named entity recognition; coreference resolution; corpus annotation; parsing and morphology; text summarisation and simplification; event extraction; fact checking and rumour analysis; NLP for healthcare; and NLP for social media.

We would like to thank all members of the Programme Committee and all additional reviewers. Together they have ensured that the best papers were included in the proceedings and have provided invaluable comments for the authors.

Finally, special thanks go to the University of Wolverhampton, the Bulgarian Academy of Sciences, the AComIn European project, and Ontotext for their generous support of RANLP.

Welcome to Varna and we hope that you enjoy the conference!

The RANLP 2017 Organisers

Automatic Identification of AltLexes using Monolingual Parallel Corpora

Elnaz Davoodi and **Leila Kosseim**
Department of Computer Science and Software Engineering
Concordia University
1515 Ste-Catherine Blvd. West
Montreal, Quebec, Canada, H3G 2W1
`ed.davoodi@gmail.com`
`leila.kosseim@concordia.ca`

Abstract

The automatic identification of discourse relations is still a challenging task in natural language processing. Discourse connectives, such as *since* or *but*, are the most informative cues to identify explicit relations; however discourse parsers typically use a closed inventory of such connectives. As a result, discourse relations signaled by markers outside these inventories (i.e. AltLexes) are not detected as effectively. In this paper, we propose a novel method to leverage parallel corpora in text simplification and lexical resources to automatically identify alternative lexicalizations that signal discourse relation. When applied to the Simple Wikipedia and Newsela corpora along with WordNet and the PPDB, the method allowed the automatic discovery of 91 AltLexes.

1 Introduction

Understanding a text goes beyond understanding its textual units in isolation; the relation between these units must also be understood. Discourse connectives such as *since*, *but*, etc. are often used to explicitly connect textual units and signal the presence of specific *explicit* discourse relations such as CONTRAST, CAUSE, etc. The Penn Discourse Tree Bank (PDTB) framework (Prasad et al., 2008) defined a closed list of discourse connectives. *Implicit*, *AltLex* and *EntRel* are other realizations of discourse relations in the PDTB. AltLex (or alternative lexicalization relations), which are understudied in computational discourse processing, are signalled using an open list of lexical markers that are not part of the PDTB inventory of discourse connectives.

Figure 1 shows a pair of sentences that convey the same information; however only one sentence contains a discourse connective from the PDTB inventory. Hence, a discourse parser using the PDTB inventory of connectives would easily identify the explicit CONTRAST relation in the first sentence but will likely not tag the second sentence because *whilst* is not part of the PDTB inventory of discourse connectives. However, the writer's intention can be understood using other means such as the use of an alternative lexical marker (i.e. AltLex), a change of tense, a structural signal, etc. Thus, discourse parsers can benefit from the automatic identification of AltLexes that can signal discourse relations.

According to Pitler et al. (2009), discourse connectives constitute strong clues to detect explicit relations, hence discourse parsers have relied on them as valuable features in order to identify explicit discourse relations automatically (Lin et al., 2014). Similarly, the presence of alternative lexical markers is a strong indicator of an AltLex relation; however since the list of such markers is open, identifying them is a challenge.

2 Background

Discourse connectives (Blakemore, 1987; Knott and Dale, 1994; Schourup, 1985) are the most informative signals of explicit discourse relations (Pitler et al., 2009). However, they are not well-defined in linguistics. Levinson (1983) defined discourse connectives as words and phrases such as *after all*, *actually*, etc. that connect an utterance to the prior discourse. Zwicky (1985) considered discourse connectives as a class of particles, but did not specify what particles are considered as discourse connectives. Schiffrin (1988) also defined discourse connectives as words that connect dependent textual units in a discourse. According to Schiffrin, discourse connectives do

Complex	Simple
These works he produced and published himself, **whilst** his much larger woodcuts were mostly commissioned work. [Non-Explicit]	He created and published his works himself, **but** his larger works were mostly commissioned work to be sold. [Explicit CONTRAST]

Figure. 1: No explicit relation is detected in the complex sentence (left), but an explicit CONTRAST relation is identified in the simple sentence (right). The example is taken from the Simple English Wikipedia corpus (Coster and Kauchak, 2011)

not belong to any linguistic class and except for a few discourse connectives such as *oh* and *well*, most carry meaning. Redeker (1991) revised this definition; even though she agreed that discourse connectives have meaning by themselves, she argued that they should contribute to the semantic interpretations of the discourse. Apart from research efforts aiming at defining discourse connectives, another line of research has focused on providing a list of discourse connectives in English (Prasad et al., 2008; Andersen, 2001; Blakemore, 2002; Fischer, 2000; Sanders et al., 1992; Knott, 1996) and other languages (Pasch, 2003; Travis, 2005). While most of these inventories have been built by hand, some work has attempted to identify them automatically. Laali and Kosseim (2014) used the Europal parallel corpus and collocation techniques to induce French connectives from their English counterparts. Following this work, Hidey and McKeown (2016) built a parallel corpus of causal and non-causal AltLexes using word alignment with PDTB discourse connectives as initial seeds. Our work is different from these as we use already existing parallel corpora in text simplification and extract discourse information automatically using the state of the art discourse parser. In addition, instead of focusing on only one relation, we generalize the problem to all PDTB relations. We also use external resources which are shown to have advantages over word alignment (Versley, 2010) in similar tasks. Lastly, the PDTB AltLexes only captures inter-sentence relations. Our contribution overcomes this limitation by identifying intra-sentence relations.

3 Identification of Discourse Connectives

In order to automatically identify AltLexes, we used the notion of text simplification (Siddharthan, 2014; Kauchak, 2013). While two texts may convey the same meaning, they may exhibit different complexity levels (Pitler and Nenkova, 2008; Davoodi and Kosseim, 2016). This difference in complexity level may be the result of various linguistic choices; at the lexical level (e.g. using fre-

> **Complex: When** the show was broadcast, Rupert Boneham won the million dollars. [Explicit SYNCHRONY]
> **Simple:** Rupert Boneham won the million dollars.

Figure. 2: An example of the removal of a discourse argument and consequently the removal of a discourse relation.

quent vs. abandoned words), the syntactic level (e.g. using active vs. passive voice) or even the discourse level (e.g. using an implicit vs. an explicit discourse relation). The main assumption in text simplification is that it is possible to reduce a text's complexity while preserving its meaning as much as possible. Because discourse relations are semantic in nature, we can therefore assume that they are also preserved as much as possible during text simplification. However, the lexical realization of discourse relations (i.e. explicit versus non-explicit) or the choice of a discourse connective (e.g. *but* vs. *however*) may change. The removal of a discourse relation may happen if the discourse argument is considered non-essential. For example, Figure 2 shows a pair of aligned sentences where the complex version contains an explicit SYNCHRONY relation signalled by *when*; while the discourse argument and consequently the explicit discourse connective has been removed in the simple version. Hence, given a sentence and its simplified version, three phenomena can occur:

1. a discourse connective is replaced by another (e.g. *although* ⇒ *though*),
2. a discourse connective is replaced by another lexical choice (i.e. word or phrase) which is not considered as a discourse connective in the inventory used (e.g. *although* ⇒ *despite*), or
3. a discourse connective is removed completely.

In cases (1) and (2) above, the discourse relation is preserved, while in case (3) the discourse relation is either removed or changed to an implicit relation. The focus of this paper is to study case (2) and use such a phenomenon to automatically identify AltLexes.

3.1 Data Sets

To discover AltLexes automatically, we created two sentence-aligned data sets using standard corpora in text simplification. The first data set was created from the Simple English Wikipedia corpus (Coster and Kauchak, 2011); the other was created from the Newsela corpus (Xu et al., 2015).

The Simple English Wikipedia (SEW) corpus (Coster and Kauchak, 2011) contains two sections: 1) article-aligned and 2) sentence-aligned. Here, we used the sentence-aligned section, which contains 167,686 pairs of aligned sentences.

In order not to overfit to a specific corpus, we also used the Newsela (News) corpus (Xu et al., 2015). This corpus contains 1,911 English news articles which have been manually re-written at most 5 times, each time with decreasing complexity level. We used this article-aligned corpus to align it at the sentence-level using an approach similar to (Coster and Kauchak, 2011). Then, two native English speakers evaluated the alignments. The Kappa inter-annotation agreement was 0.898 computed on 100 randomly chosen alignments.

3.2 Methodology

According to the PDTB framework, each Altlex can be substituted with at least one discourse connective (Prasad et al., 2008). Based on this, to discover AltLexes automatically, we first parsed both sides of the aligned sentences of both data sets to extract discourse information. This was done using the PDTB-style End-to-End parser (Lin et al., 2014). This parser was selected as it is currently the best performing parser to identify explicit relations, with an F-measure between 80.61% and 86.77% depending on the evaluation criteria. Because it uses the PDTB framework, the parser uses the inventory of 100 discourse connectives from the PDTB. The result of this tagging was categorized into one of the following cases:

1. *NonExp-NonExp*: a non-explicit[1] discourse relation occurs in both sentences.
2. *Exp-Exp*: the same discourse relation and discourse connective occur in both sentences.
3. *NonExp-Exp*: a non-explicit relation occurs in the complex sentence, but an explicit one is used in the simple sentence.
4. *Exp-NonExp*: an explicit relation occurs in the complex sentence, but no relation is used in the simple sentence.

Change	SEW-based DS		News-based DS	
(1) NonExp-NonExp	116,852	69.68%	18,384	50.00%
(2) Exp-Exp	19,735	11.76%	2,660	7.23%
(3) NonExp-Exp	**7,868**	**4.69%**	**1,129**	**3.07%**
(4) Exp-NonExp	**9,490**	**5.65%**	**1,733**	**4.71%**
(5) other	13,741	8.22%	12,862	34.99%
Total	167,686	100%	36,768	100%

Table 1: Frequency of the discourse changes across complexity levels in the SEW-based and News-based data sets.

5. *Other*:
   a. *Same Relation-Different Connective*: the same explicit relation is used but with different discourse connectives in both sentences.
   b. *Different Relation-Different Connective*: a different explicit relation and a different discourse connective are used.
   c. other cases including several explicit relations within a single sentence.

Table 1 shows the frequency of these transformations in the two data sets. To discover AltLexes, we only considered cases (3) and (4), where only one side of the alignment includes a PDTB discourse connective. This gave rise to a total of 20,220 aligned sentences. We then used two external resources: 1) the paraphrase database (PPDB) (Ganitkevitch et al., 2013) and 2) Word-Net (Miller, 1995). The PPDB comes in six sizes from S to $XXXL$. The smaller versions of the PPDB contain more precise paraphrases with higher confidence scores; while the larger versions have more coverage. We choose the PPDB version L to have a good compromise between the precision and coverage of paraphrases. We took the discourse connective from the explicit side and looked for an alternative lexicalization (a synonym or paraphrase) in the external resources. If any of its alternative lexicalization appeared in the non-explicit side, we considered it as an AltLex to signal the relation. We then replaced the AltLex with the discourse connective from the explicit side and parsed the new sentence with the PDTB-style End-to-End parser again. This process is shown in Figure 5. On average, each discourse connective was replaced by 23.2 alternative lexicalizations taken from the PPDB and 12.3 from WordNet.

If the parser detected the same relation (see Figure 5), then the potential marker was considered as an AltLex. On the other hand, because the End-to-End discourse parser uses both the discourse connective and syntactic features, if it was not capable

---

[1]A non-explicit discourse relation can refer to an implicit, and AltLex discourse relation or to no discourse relation.

Discourse Relation	SEW-based Data set			News-based Data set			Overall	
	Alignments	New AltLexes from PPDB	WordNet	Alignments	New AltLexes from PPDB	WordNet	Alignments	New markers from SEW ∪ News
ASYNCHRONOUS	2,561	15	3	327	6	2	2,888	20
SYNCHRONY	1,990	2	1	395	0	0	2,385	3
CAUSE	1,359	18	1	256	3	0	1,615	19
CONDITION	296	0	0	141	1	0	437	1
CONTRAST	2,568	6	1	667	5	6	3,235	9
CONCESSION	393	3	0	64	0	0	457	3
CONJUNCTION	7,738	25	1	914	12	3	8,652	27
INSTANTIATION	159	3	1	33	0	0	192	3
RESTATEMENT	63	1	0	13	0	0	76	1
ALTERNATIVE	220	5	0	51	1	0	271	5
EXCEPTION	8	0	0	0	0	0	8	0
LIST	3	0	0	1	0	0	4	0
Total	17,358	79	8	2,862	28	11	20,220	91

Table 2: Number of *Exp-NonExp* and *NonExp-Exp* alignments and newly identified AltLex types.

Complex	Simple
Now they have drones in 15 states, including California and Texas. **Before** they started the business, the two covered fields on foot or in vehicles. [Explicit ASYNCHRONOUS]	Now they have drones in 15 states, including California and Texas. Fiene used to check farm fields on foot or with vehicles.

Figure. 3: An example of discourse marker and potential AltLex having different syntactic class across complexity levels.

of detecting the relation in the replaced sentences, we concluded that either (1) the relation existed, but the parser could not detect it, (2) the AltLex does not signal the discourse relation or (3) the relation does not exist (see Figure 4). Because we did not use any syntactic filter, the replacement of the discourse connective may alter the syntax of the sentence such that the parser is unable to detect the relation. This is why, regardless of the reason, if the parser was not able detect the relation, we discarded the AltLex.

4 Results and Discussion

Table 2 shows the number of sentence alignments mined and the number of potential AltLexes (i.e. type count) identified in each data set for each level 2 PDTB relation. Overall, by mining 17,358 NonExp-Exp and Exp-NonExp alignments, the SEW-based data set allowed the discovery of 79 AltLexes from the PPDB and 8 from WordNet; whereas, the News-based data set, providing only 2,862 alignments, allowed the discovery of 28 AltLexes from PPDB and 11 from WordNet. Using both corpora and both lexical resources, the method found 91 AltLexe tokens, which account for 533 AltLexes. It is interesting to note that, overall, the approach did not find any alternate lexicalizations for some relations such as LIST or EXCEPTION and only one for CONDITION. It is not clear if this is because these relations are typically signalled using a rather fixed inventory of discourse markers or because of the low number of such alignments. Indeed, in the PDTB, out of

Complex: It's a very special place **because** this site, this area, has been tied to the history and life of African-Americans since about the early 1800s. [Explicit CAUSE]
Simple: It has been tied to the history and life of African-Americans *since* [SYNONYM OF BECAUSE] about the early 1800s.
⇓
Simple after substitution: It has been tied to the history and life of African-Americans **because** about the early 1800s.

Figure. 4: An example which does not lead to an AltLex. After the substitution, the CAUSE relation is still not identified.

Complex: Today, the comic arm of the company flourishes *despite* [SYNONYM OF THOUGH] no longer having its own universe of super powered characters.
Simple: Today, the company does very well even **though** they do not have their own universe of super powered characters. [Explicit CONTRAST]
⇓
Complex after substitution: Today, the comic arm of the company flourishes **though** no longer having its own universe of super powered characters. [Explicit CONTRAST]

Figure. 5: An example which leads to an AltLex. After the substitution, the CONTRAST relation is identified.

624 tagged AltLex relations, only 6 are labeled as RESTATEMENT, 1 as EXCEPTION and 2 as CONDITION.

On the other hand, relations such as CONJUNCTION, ASYNCHRONOUS and CAUSE provided a large number of alignments from which we identified a variety of AltLexes. For example, the PPDB identified "*caused by*", "*resulting*", "*causing*", "*this being so*", etc. as AltLexes to signal a CAUSE relation. The full inventory of the au-

tomatically identified AltLexes can be found at: http://Anynomous.

In addition, as can be seen in Table 2, the number of potential AltLexes coming from the PPDB is higher than the number of AltLexes coming from WordNet. This may be due to the difference of the coverage of these two resources as WordNet is smaller than the PPDB. Another possible reason is that each word in PPDB has a list of paraphrases with various syntactical classes. Thus, if the syntactic class of a discourse marker is changed in the simplification process, it is more probable that the PPDB covers more syntactical variations of the discourse marker compared to WordNet. Figure 3 shows an example taken from the Newsela corpus. In this example, the discourse marker *before* in the complex version signals ASYNCHRONOUS relation, but is tagged as subordinating conjunction. In the paraphrase database, *used to* is one of the paraphrases of discourse marker *before*. In the simple version of this example, the verb *used to* is signalling the same relation (i.e. ASYNCHRONOUS relation) which is captured as an AltLex.

5 Future Work

As future research, we plan to assign a confidence level to the automatically identified AltLexes by using a syntactic filter to replace potential markers only if they lead to syntactically correct sentences and by using their frequency of appearance in the parallel corpora. Another interesting line of research would be to evaluate if discourse parsers can increase their performance using such new list.

Acknowledgement

References

Gisle Andersen. 2001. *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents*, volume 84. John Benjamins Publishing.

Diane Blakemore. 1987. *Semantic constraints on relevance*. Oxford.

Diane Blakemore. 2002. *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*, volume 99. Cambridge University Press.

William Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*. Association for Computational Linguistics, pages 1–9.

Elnaz Davoodi and Leila Kosseim. 2016. On the contribution of discourse structure on text complexity assessment. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles, pages 166–174.

Kerstin Fischer. 2000. *From cognitive semantics to lexical pragmatics: The functional polysemy of discourse particles*. Walter de Gruyter.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*. Association for Computational Linguistics, Atlanta, Georgia, pages 758–764. http://cs.jhu.edu/ ccb/publications/ppdb.pdf.

Christopher Hidey and Kathleen McKeown. 2016. Identifying causal relations using parallel wikipedia articles.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *ACL (1)*. pages 1537–1546.

Alistair Knott. 1996. A data-driven methodology for motivating a set of coherence relations .

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes* 18(1):35–62.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*. pages 610–619.

Stephen C Levinson. 1983. *Pragmatics (Cambridge textbooks in linguistics)*. Cambridge University Press.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20(2):151–184.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Renate Pasch. 2003. *Handbuch der deutschen Konnektoren: linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*, volume 1. Walter de Gruyter.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Association for Computational Linguistics, pages 683–691.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 186–195.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie L. Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakesh, Morocco.

Gisela Redeker. 1991. Linguistic markers of discourse structure. *Linguistics* 29(6):1139–1172.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes* 15(1):1–35.

Deborah Schiffrin. 1988. *Discourse markers*. 5. Cambridge University Press.

Lawrence Clifford Schourup. 1985. *Common discourse particles in English conversation*. Dissertations-G.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165(2):259–298.

Catherine E Travis. 2005. *Discourse markers in Colombian Spanish: A study in polysemy*, volume 27. Walter de Gruyter.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. pages 83–82.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics* 3:283–297.

Arnold M Zwicky. 1985. Clitics and particles. *Language* pages 283–305.