Toward Cross-theory Discourse Relation Annotation

Peter Bourgonje and **Olha Zolotarenko**
Applied Computational Linguistics
University of Potsdam / Germany
firstname.lastname@uni-potsdam.de

Abstract

In this exploratory study, we attempt to automatically induce PDTB-style relations from RST trees. We work with a German corpus of news commentary articles, annotated for RST trees and explicit PDTB-style relations and we focus on inducing the implicit relations in an automated way. Preliminary results look promising as a high-precision (but low-recall) way of finding implicit relations where no shallow structure is annotated at all, but mapping proves more difficult in cases where EDUs and relation arguments overlap, yet do not seem to signal the same relation.

1 Introduction

The task of *discourse processing* or *discourse parsing* refers to the extraction of coherence relations between abstract entities (propositions, etc.) from plain text. Within this field, three of the most popular frameworks in terms of influence and available annotated data; the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003), each have their own characteristics when it comes to representing these coherence relations, both at elementary (segmentation) level, internal structure (global vs. local) and in terms of sense sets used. Generating annotated data for discourse parsing is a costly process (as reflected by the relatively small size of available corpora and the low inter-annotator agreement figures ((Carlson et al., 2001), (Asher et al., 2016))), and available corpora as a result are relatively small compared to corpora annotated for other NLP tasks. Enabling annotations from one framework to enrich annotations in another thus seems a fruitful goal to pursue. For at least two such corpora, annotations on the same source text for two different frameworks exist; the PDTB and the RST-Discourse Treebank (RST-DT) both use (an overlapping set of) (English) Wall Street Journal articles, and the Potsdam Commentary Corpus has PDTB-style annotations and RST annotations on a set of (German) news commentary articles.

We are working with the Potsdam Commentary Corpus. As a first step toward comparing the relations in both frameworks in independently annotated text, we attempt to map the segments of both frameworks. The main contribution of this paper is to investigate the feasibility of enriching a shallow, PDTB-style annotation layer by exploiting RST-trees for the same text. An overview of similar approaches is listed in Section 2, the corpus we work with is described in Section 3. Results of aligning segments and relations are presented in Section 4 and a brief wrap-up is provided in Section 5.

2 Related Work

There is a large amount of literature on both the RST and PDTB frameworks, but we focus here on the mapping between the two. Earlier work on the same corpus is described in Scheffler and Stede (2016), where PDTB relations are projected onto RST relations (the opposite of what we are doing in this paper) to obtain an overview of sense synergies. The authors note that of the 2,536 RST relations in the corpus, only 932 were marked by an explicit connective, rendering the majority (63%) implicit, which is a promising percentage given our goal of enriching the shallow layer with implicit relations (see Section 3 for more details). Several attempts have been made at unifying the set of senses used in the difference discourse relation frameworks, but most of them do so from a theoretical perspective, i.e. Rehbein et al. (2016), Benamara and Taboada (2015), Bunt and Prasad

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 7–11
Minneapolis, MN, June 6, 2019.

(2016), Chiarcos (2014) and Sanders et al. (2018).

A notable exception is the practical approach based on the PDTB and the RST-DT described by Demberg et al. (2017). The PDTB (Prasad et al., 2008) is annotated on the same set as the RST-DT (Carlson et al., 2002), but the former is considerably larger, with over 1.3m tokens compared to ca. 200k tokens, respectively. This makes the exploitation of shallow annotations to construct RST-trees a potentially more promising (yet probably more complex) venture. Our data however is already annotated for RST-trees and only partly annotated on a shallow level, and also in German (as opposed to English for the PDTB and RST-DT). The general aim of bringing together different discourse frameworks is at the heart of the 2019 DISRPT workshop[1] and hopefully the workshop will inspire more work in this direction.

3 Data & Method

The corpus under investigation is the Potsdam Commentary Corpus (PCC) (Stede and Neumann, 2014), a German collection of news commentary articles from a local German newspaper containing ca. 33k words. The RST layer has been annotated according to the structural constraints defined by Mann and Thompson (1988), using a slightly modified relation set and relations with centrally embedded segments are not annotated in the corpus. The entire corpus contains 176 RST trees (for the 176 articles), containing 3,018 Elementary Discourse Units (EDUs). The shallow (PDTB-style) layer has been annotated only for relations using an explicit connective (using the definition of Pasch et al. (2003)). An explicit relation comprises the connective token(s), the external argument (*arg1*) and the internal argument (*arg2*). There are 1,110 explicit relations in the corpus, meaning that we have twice that number (2,220) of arguments. Both layers have been annotated independently from each other. For further details on annotation procedures, we refer to Stede and Neumann (2014).

Before proceeding with our mapping procedure, it is important to note that the nature of the segments (EDUs in the RST layer, arguments in the shallow layer) are by design of a different type. While in the RST approach, segmentation is a first and essential step in annotating or analysing a text, this is not the case in the PDTB approach. Instead, the latter first identifies explicit connectives and then locates arguments according to the "minimality principle", which prescribes that only as much material should be included in the argument as is *minimally required* to interpret the relation. Arguments of explicit relations and RST EDUs will be the types of segments we are comparing. Arguments for explicit and implicit relations are of a fundamentally different type (with implicit relation arguments being typically entire sentences, or complete clauses delimited by a (semi-)colon (see Prasad et al. (2017) for more details). However, since we do not have implicit relations in our corpus (in fact, this is exactly what we intend to infer from the RST relations), we can discard this difference during the mapping phase. Section 4 will include more details on the implications of this discrepancy for induced relations.

Additionally, in the RST layer we expect to find many more relations than in the shallow layer. Not only because implicit relations are not included in the latter, but also because RST, in contrast to the shallow approach, includes complex relations, i.e. relations where one or both of the components can be complex units. Because we intend to extract shallow relations, we discard all complex RST relations. The relation between segment 17 and 18 in Figure 1 is taken into account, but the relation involving segment 16 (the conjunction relations 16-18) is not, since one of its nodes is a complex node.

Demberg et al. (2017) implement a more complex, and more complete mapping algorithm, incorporating the Strong Nuclearity hypothesis (Marcu, 2000), which would result in more RST relations (since we could then also consider the relation between a "flat" nucleus and that of a complex structure). Due to the exploratory nature of our approach, we leave this to future work. Our filtering thus results in 2,111 non-complex RST relations in the corpus, compared to the 1,110 relations in the shallow layer. Recall that we have 3,018 EDUs in the RST layer and 2,220 arguments in the shallow layer. Looking at a very general characteristic, the token length, EDUs and *arg1* and *arg2*[2] segments seem relatively comparable. The average length (in tokens) and the standard deviation for the EDUs, *arg1* and *arg2* segments respectively are 11.0/6.1, 13.5/7.3 and 13.0/10.4.

When attempting to map relations, we start

[1] https://sites.google.com/view/disrpt2019

[2] Connective tokens are included in *arg2*.

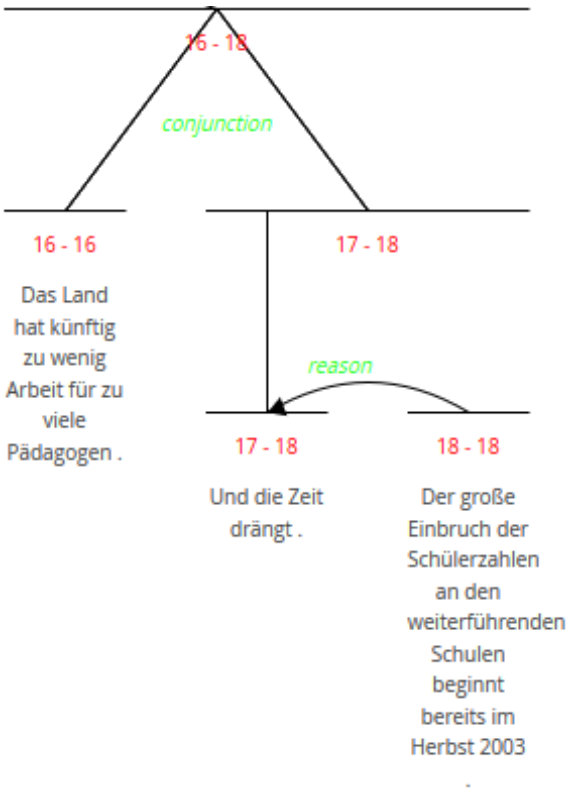

Figure 1: Part of an RST tree

from the RST relation and distinguish three different scenarios:

- There is a complete match, given a small tolerance[3], for the two EDUs and the *arg1* and *arg2*.
- One of the EDUs matches one of the arguments, but the other argument does not match the other EDU(s).
- There is no overlap between the EDU and any argument of any relation.

For the 2,111[4] RST relations, we find 305 complete matches. The second category (where one of the EDUs matches one of the arguments) contains 323 cases, leaving 1,483 cases for the last category. At this point, we leave the further categorisation and investigation of the 323 cases where one EDU matches to future efforts, because the ways in which an existing (explicit) relation interacts with a potential implicit relation induced from the RST layer need careful investigation first. It could be the case that the annotations on both levels refer to the same coherence relation in the text but the arguments are annotated differently. Or they may describe a different relation (as is the case in Example (1) below). We first turn to the remaining 1,483 cases, as these are likely to provide the best candidates for (semi-)automatically adding the RST relations to the shallow layer as implicit relations.

4 Analysis & Results

We manually checked the outcome of the mapping process for 17 documents (ca. 10% of the entire corpus). In these 17 documents, we found 64 RST relations of the third type, i.e. relations for which there was no overlap between the EDU and any argument of any relation (given our tolerance of 75%). Focusing on these cases, we still find many cases (21) where there is partial overlap (but below our threshold) and segmentation differs. An example is shown in (1), where the *arg1* and *arg2* are marked in italics and bold face, respectively. The two EDUs that were recognised in the RST layer however, were "Nun wird der Katastrophenschutz einen neuen Stellenwert bekommen." (*Now disaster prevention will take on a new significance.*) and "Der Landkreis und die Kommunen, vordergründig bedroht oder einfach nur in verständlicher Sorge, sind auf Hilfe angewiesen." (*The administrative district and the municipalities, ostensibly threathened or simply with understandable concern, are dependent on help.*)

(1) "Nun wird der Katastrophenschutz einen neuen Stellenwert bekommen. Der Landkreis und die Kommunen, *vordergründig bedroht* oder **einfach nur in verständlicher Sorge**, sind auf Hilfe angewiesen."

Now disaster prevention will take on a new significance. The administrative district and the municipalities, ostensibly threatened *or* ***simply with understandable concern****, are dependent on help.*

Before unification at the segmentation level is realised, these cases are difficult to process, as both annotation layers essentially talk about different propositions.

[3] When comparing EDUs and arguments, we assume two segments to match when there is a >75% token overlap, to include cases where the difference is just a punctuation symbol or function word.

[4] Note that this number is smaller than the 2,536 mentioned in Scheffler and Stede (2016) because we use non-complex relations only, also resulting in fewer complete matches (their 452 compared to our 305).

There were several cases where one *arg1* or *arg2* contained two EDUs, meaning that the RST layer made a more fine-grained distinction. This was the case for 7 *arg1*s and 9 *arg2*s. An example is shown in (2), which contains an *arg2* in the PDTB layer (i.e. the first two tokens ("Und so" *And so*) are the connective, and the remaining "muss Landrat ... Folgen angeht." (*district administrator ... its consequences.*) is the entire *arg2*). This argument contains two EDUs: "Und so muss Landrat Christian Gilde jetzt eine gewisse Hilflosigkeit erkennen lassen," (*And so district administrator Christian Gilde must now admit a certain helplessness,*) and "was das Reagieren auf möglichen Terror und seine Folgen angeht." (*when it comes to reacting to possible terror and its consequences.*).

(2) "Und so **muss Landrat Christian Gilde jetzt eine gewisse Hilflosigkeit erkennen lassen, was das Reagieren auf möglichen Terror und seine Folgen angeht.**"

And so ***district administrator Christian Gilde must now admit a certain helplessness when it comes to reacting to possible terror and its consequences.***

Example (2) is a good candidate for enriching the shallow layer, as it is introducing structure (an implicit relation) inside an entire argument in the PDTB layer.

This leaves 27 cases where there was no annotation in the PDTB layer at all, marking these as good candidates for (semi-)automated addition as implicit arguments as well. The distribution of senses is quite diverse, with 6 cases annotated (in the RST tree) as e-elaboration, 6 as joint, 5 as span, 3 as sequence and the remaining distributed over conjunction, evaluation-s, contrast, list, elaboration, purpose and reason. Earlier work on sense unification from Scheffler and Stede (2016) can guide in automatically assigning a PDTB sense for these cases. An important note is that there is a fundamental difference between the arguments of explicit and that of implicit relations, as mentioned earlier in Section 3. The arguments of implicit relations typically are sentences and the average sentence length and standard deviation in the PCC is 15.2/8.9 respectively, compared to 11.0/6.1 for EDUs. Using EDUs to populate implicit relations may result in a skewed distribution of implicit arguments. Especially if this semi-automatic step is done first, and then the blanks are filled out by annotating implicit relation in a manner similar to the PDTB one. Arguably, the RST segmentation is more meaningful than the segmentation procedure for implicit PDTB relation stipulation (which links sentences without any further consideration). One way to proceed, after this first semi-automatic step, could therefore be to start out with EDUs from the RST layer and assign them implicit relations if they are not involved in an explicit relation. This effectively puts the segmentation task central to shallow annotations as well, which deviates from the original annotation strategy for shallow discourse relations. As mentioned above, our use case may be somewhat unusual (with the more complex, expensive-to-obtain RST trees available, but only the explicit part of the shallow relations), but first steps indicate that this first phase of our approach is essentially a high-precision, but relatively low-recall means of (semi-)automatically finding implicit relations.

5 Conclusions & Outlook

We explore the feasibility of exploiting discourse annotations following the RST framework to add implicit relations in PDTB-style for a German corpus of news commentary articles annotated for explicit discourse relations (in PDTB-style) only. Our use case may be non-typical, with RST annotations typically being harder and more costly to obtain than shallow PDTB-style annotations, but the first results for adding implicit relations in a semi-automated way look promising. Several issues need more detailed analysis though. Partially overlapping relations (where one of the EDUs matched with one of the arguments) can be about wholly different relations (hence must not be mapped without further investigation), and we focus first on pieces of text for which no PDTB-style annotation exists at all. We consider flat, non-complex RST relations only and our approach can be improved by using the Strong Nuclearity Principle as applied in earlier work on mapping PDTB and RST relations. Segmentation differences between EDUs and implicit relation arguments specifically need more investigation, and generally arriving at a (theory-neutral) standard for discourse segmentation may prove to be very beneficial for the purpose of cross-theory annotation augmentation.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 323949969. We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this manuscript.

References

N. Asher, A. Lascarides, S. Bird, B. Boguraev, D. Hindle, M. Kay, D. McDonald, and H. Uszkoreit. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *LREC*. European Language Resources Association (ELRA).

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152. Association for Computational Linguistics.

Harry Bunt and R. Prasad. 2016. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Proceedings 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 45–54.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank, ldc2002t07.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Vera Demberg, Fatemeh Torabi Asr, and Merel Scholman. 2017. How consistent are our discourse annotations? insights from mapping RST-DT and PDTB annotations. *CoRR*, abs/1704.08893.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8:243–281.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA.

Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Herrmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *In Proceedings of LREC*.

Rashmi Prasad, Katherine Forbes-Riley, and Alan Lee. 2017. Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the pdtb. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16. Association for Computational Linguistics.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *LREC*.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 0(0). Exported from https://app.dimensions.ai on 2019/02/06.

Tatjana Scheffler and Manfred Stede. 2016. Mapping pdtb-style connective annotation to RST-style discourse annotation. In *Proceedings of KONVENS*, Bochum, Germany.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).