The utility of discourse parsing features for predicting argumentation structure

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, Manfred Stede
University of Potsdam, Germany
{hewett,rane,harlacher,stede@uni-potsdam.de}

Abstract

Research on argumentation mining from text has frequently discussed relationships to discourse parsing, but few empirical results are available so far. One corpus that has been annotated in parallel for argumentation structure and for discourse structure (RST, SDRT) are the 'argumentative microtexts' (Peldszus and Stede, 2016a). While results on perusing the gold RST annotations for predicting argumentation have been published (Peldszus and Stede, 2016b), the step to automatic discourse parsing has not yet been taken. In this paper, we run various discourse parsers (RST, PDTB) on the corpus, compare their results to the gold annotations (for RST) and then assess the contribution of automatically-derived discourse features for argumentation parsing. After reproducing the state-of-the-art Evidence Graph model from Afantenos et al. (2018) for the microtexts, we find that PDTB features can indeed improve its performance.

1 Introduction

The argumentative structure of texts, as captured, for instance, by schemata from Peldszus and Stede (2013) or Stab and Gurevych (2014), is represented by tree structures that suggest a certain similarity to accounts of discourse structure, such as in Rhetorical Structure Theory (Mann and Thompson, 1988) or Segmented Discourse Representation Theory (Asher and Lascarides, 2003). These approaches aim at accounting for the *coherence* of texts, which is clearly related – though not identical – to the structure of complex arguments. This is not a new observation (Habernal and Gurevych, 2017), but we are not aware of many empirically-grounded studies of the correspondences between the two realms. A corpus that facilitates such experiments is the 'argumentative microtext corpus' (Peldszus and Stede, 2016a), as it offers annotation not only for argumentation but also for discourse structure in terms of RST and SDRT. While there is evidence that RST trees can "in principle" be helpful for parsing the argumentation (based on the gold annotations; see Peldszus and Stede, 2016b), we are not aware of experiments which try to verify such effects with automatic parsers. Our work aims to bridge this gap. We use common parsers for RST and for Shallow Discourse Parsing (specifically the Penn Discourse Treebank, henceforth PDTB), run them on the microtexts, and first compare the RST output to the gold annotations, in order to assess the prospects of the idea. Having selected the most promising parsers, we then compute a set of features from their output and add them to a state-of-the-art implementation of argumentation parsing on the microtexts (Afantenos et al., 2018). The results indicate that the parsed PDTB features do in fact improve the accuracy of the argumentation annotation.

Section 2 discusses related work, and Section 3 describes the corpus and the discourse parsers we used. Initial analyses of parser results are given in Section 4, and the experiments on predicting argumentation structure are reported in Section 5. The paper closes with some conclusions in Section 6.

2 Related work

A number of researchers have studied connections between discourse structure and argumentation. Cabrio et al. (2013) look at the link between PDTB relations and the argumentation schemes from Walton et al. (2008). They find, for example, that the PDTB relation 'expansion' corresponds to the 'Argument by Example', which can be defined as when the second argument offers a summary or a conclusion based on the first argument. Generally, the presence of connectives or other discourse markers has often been employed

Proceedings of the 6th Workshop on Argument Mining, pages 98–103
Florence, Italy, August 1, 2019.

for detecting argument components and relations between them. Stab and Gurevych (2014) compile a list of 55 discourse markers which indicate argumentative discourse and use these as features to detect the argumentative role in German essays. Eckle-Kohler et al. (2015) instead look at German news items which are annotated with the argumentative roles 'claim' and 'premise' (with various sub-categories). They found that both single discourse markers and semantic groups of such markers occurred in significant correlation with claims or premises. Turning specifically to RST, Green (2010) proposes the ArgRST annotation scheme, which represents both argumentation and discourse analysis in the same structure. *Inter alia*, she finds parallels between the RST relation 'evidence' and the premise and claim of an argument.

Finally, Peldszus and Stede (2016b) present a qualitative study on the mapping from manual RST annotations to argumentation structure and also conduct experiments using a new feature set which is based exclusively on the gold RST annotation (of the 'microtext' corpus; see Section 3.1). These features include the position of the segment in the text, whether a segment has incoming or outgoing edges, and the type of RST relation between segments, amongst others. They showed that especially two subtasks of argumentation structure parsing in the microtexts (finding the central claim and the attachment point of segments) can clearly benefit from these features. Our project is a continuation of that study, as we essentially replicate the model, but use automatically parsed RST trees instead of the gold annotations, in order to assess a "real world" scenario.

3 Data and parsers

3.1 Argumentative microtexts

Part 1 of the microtexts corpus (Peldszus and Stede, 2016a) is a freely available[1] parallel corpus of 112 short texts with 576 argumentative segments. They were originally written in German and have been professionally translated to English, preserving the segmentation where possible. The texts have been collected in a controlled text generation experiment using a short instruction. All texts have been annotated with argumentation structure according to the scheme of Peldszus and Stede (2013), i.e., trees with one claim and support/attack relations between the segments. Furthermore, various other layers of annotation have been produced, including RST trees (Stede et al., 2016). Later, Musi et al. (2018) conducted a study comparing the RST trees to annotations of argumentation schemes.

[1] http://angcl.ling.uni-potsdam.de/resources/argmicro.html

3.2 Argumentation parsing

Various researchers have used slightly different approaches to automatically parse the argumentation structure in the microtexts. Peldszus and Stede (2015) decompose the problem into the four subtasks of finding the central claim (*cc*) segment, and for each other segment its role (*ro*: proponent or opponent), its function (*fu*: support, rebut, undercut), and the segment it attaches to (*at*). They use a minimum spanning tree (MST) decoder on a so-called 'evidence graph' that combines the probabilities computed for the four subtasks. Stab and Gurevych (2016) achieved slightly better results for some subtasks using Integer Linear Programming. Potash et al. (2017) use a bidirectional LSTM encoder and achieve competitive results on the microtexts, but they solve only part of the problem (no support/attack distinction). Finally, Afantenos et al. (2018) compare ILP and MST by training a classifier for each subtask (*cc, ro, fu, at*) and use this combined distribution as input to the decoders. Their best model is a replication of the evidence graph model with MST decoding from Peldszus and Stede (2015) with some additional features, including discourse connectives for English. As this is this the model with best results for the complete problem, we will replicate it for our experiments.

3.3 Discourse parsing: first observations

We parsed a subset of the corpus with various parsers (Ji and Eisenstein, 2014; Feng and Hirst, 2014; Lin et al., 2014; Biran and McKeown, 2015), and after a manual analysis of the results, chose the systems of Feng and Hirst (2014) and Lin et al. (2014). These were used "out of the box", without having been trained on our data, to produce the automatic RST- and PDTB-parses for our study in a domain-independent way.

In a small pilot study, we compared the RST parser output to the gold argumentation structures for 10 texts of the corpus. We observed that the parser sometimes produced different segmentations, either combining segments, or using completely new boundaries. We also noted that the central claims matched the most-nuclear RST seg-

ment (for an explanation, see Section 4.1 below) in 50% of the graphs, and that 26 RST edges – out of 40 – corresponded to ARG edges. In these cases the relation labels were also coherent. For instance, the ARG relation *undercut* matched with the RST relation *concession* and *antithesis*, *support* corresponded with RST edges *explanation* and *cause*.

Likewise, for the 10 texts we checked the output of the PDTB parser and observed that again, the boundaries did not match in most cases. There were very few argumentation pairs that matched to the ARG edges, and the parser in general did not pick up on many relations, in particular implicit relations.

Due to the segment boundary mismatches we observed, we decided to use common pre-segmented text, taken from the gold-annotated corpus, as input to the parsers for all the following experiments. While this is in line with practices in related research, it has to be noted as a certain simplification of the "real world" scenario, as discourse- and argumentation parsing are not quite used out of the box anymore.

4 Quantitative analysis of parser output

In the next step, we turned to the full corpus of 112 texts. For quantitatively comparing our automatically-parsed texts to the gold-standard argumentative annotations of the microtexts, we first converted the tree structures to a dependency format, adapting the techniques described in Stede et al. (2016). These include converting multinuclear RST relations such as *joint* or *contrast* to nested binary relations by combining the sources of the relations. In a similar vein, *join* nodes in the ARG trees were converted to a *joint* edge between the two relevant segments, and *undercut* edges which target a relation between two edges were converted to target the source of the attacked relation. The PDTB parser output included relation predictions both within and across our pre-determined segments; for the purposes of this comparison we only considered the inter-segmental relations.

4.1 Central claims

The "most nuclear" (MN) segment in the RST structure can be identified by tracing down from the root node to the nucleus at each level, until reaching the lowest level (Marcu, 2000). We interpreted this for our RST trees by defining the MN as the segment or group with no parent. If it is a group, the RST tree can have more than one MN. If the ARG CC matches any of these MNs then it counts as match. There were a total of 67 matches, which represents about 60% of the corpus. The corresponding figure for gold RST and ARG from (Peldszus and Stede, 2016b) is 85%. Considering there are 5 segments in each text on average, we see the automatic result as a quite promising score.

4.2 Common undirected edges

	reb	join	sup	und	link	exa	NONE
elaboration	22	23	88	6	4	3	115
same-unit	2	0	1	0	0	0	8
joint	2	13	1	1	10	0	32
contrast	7	2	3	28	0	0	19
temporal	0	0	0	0	0	0	1
evaluation	3	0	3	0	0	0	7
summary	0	0	0	0	0	0	1
explanation	1	1	8	1	0	1	7
cause	0	3	8	1	1	0	3
topic-comment	0	0	0	0	0	0	1
background	0	9	14	0	0	0	7
attribution	0	4	0	0	0	0	3
condition	0	14	0	0	0	0	0
enablement	0	1	1	0	0	0	0
manner-means	0	1	1	0	0	0	0
comparison	0	0	2	0	0	0	0
NONE	65	10	114	23	6	4	0

Table 1: Co-occurrence matrix of the RST (rows) and ARG (columns) relations of the matching edges in the converted annotations

	join	und	reb	sup	link	exa	NONE
Temporal.Synchrony	6	1	0	10	0	0	1
Expansion.Conjunction	3	0	2	0	2	0	24
Comparison.Contrast	0	21	5	0	0	0	18
Expansion.Alternative	0	0	1	0	0	0	0
Contingency.Cause	0	0	0	9	0	0	7
Expansion.Instantiation	0	0	0	0	0	1	0
Contingency.Condition	5	0	0	2	0	0	2
Temporal.Asynchronous	0	1	0	1	0	0	0
Comparison.Concession	1	0	2	0	0	0	1
NONE	61	25	71	189	10	2	0

Table 2: Co-occurrence matrix of the PDTB (rows) and ARG (columns) relations of the matching edges in the converted annotations

RST & ARG: Although a large amount of edges in one annotation had no corresponding edge in the other annotation, there are some similarities. *Contrast* maps to *undercut* 28 times, and *elaboration* is frequently mapped to *join*, *support*, which seems plausible, and *rebuttal* which seems less so.

PDTB & ARG: Although few edges matched (73), this is in part due to the fact that only a total of 176 PDTB relations were identified by the parser, in comparison to 547 relations in the ARG

annotation. *Comparison.Contrast* maps to *undercut* 21 times and *Temporal.Synchrony* often maps to *join*, both of which seem to be a suitable mapping.

4.3 RST gold vs. parser

Besides comparing RST parses to argumentative structures, we were also interested in evaluating the RST parser on the microtexts, i.e., on their gold RST trees. To this end, we converted the gold annotations to a comparable format, which involved converting the 'span' relations (which were not present in the parser's output), adjusting the segment IDs so that they were in ascending order, and converting the more fine-grained relations to the smaller set used by the parser (using the taxonomy in Das and Taboada, 2014). We adapted the metrics to evaluate the parser output from those proposed by Joty et al. (2015); our results are given in Table 3.

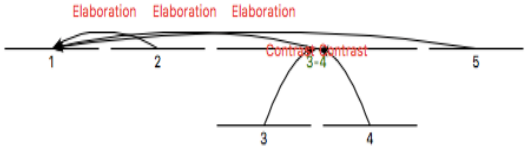
	Span	Nuclearity	Relation
F1	0.338	0.264	0.115

Table 3: RST parser evaluation, with the categories used by Joty et al. (2015) and others.

PDTB parser output

1:[Intelligence services must urgently be regulated more tightly by parliament;] 2:[this should be clear to everyone after the disclosures of Edward Snowden.] 3:[Granted, those concern primarily the British and American intelligence services,]$^{\text{Comparison.Contrast}}$ *4:[but the German services evidently do collaborate with them closely.] 5:[Their tools, data and expertise have been used to keep us under surveillance for a long time.]*

RST parser output



Best performing ARG model output

[2, 1, 'join'], [3, 1, 'rebut'], [4, 3, 'undercut'], [5, 1, 'support']

Gold ARG annotation

[2, 1, 'support'], [3, 2, 'undercut'], [4, 3, 'undercut'], [5, 4, 'support']

Figure 1: Parser and model output for microtext b005. The numbers refer to the segments. RST tree created using RSTTool (O'Donnell, 2000).

5 Prediction experiments and results

Finally, we address the task of predicting the ARG structure with the help of discourse parser output. We extended the system of Afantenos et al. (2018) and started from the feature set used by Peldszus and Stede (2016b); our own new features, listed in Table 4, will be referred to as 'RST+' and 'PDTB' respectively. The task now is to assess their contribution in comparison to the 'Default' and 'RST' features from Peldszus and Stede (2016b) and to the best performing lexical, syntactic, semantic and discourse features used by Afantenos et al. (2018). In Table 5, which shows our results, the latter are labelled as '2018'.

We experimented with different combinations of the features on two different settings of the model: the simple relation set (*support* and *attack*); and the more fine-grained full relation set (*support, example, join, link, undercut* and *rebuttal*). We used the same train-test splits as in Peldszus and Stede (2015), which involved 10 iterations of 5-fold cross validation. The results for the full relation set were marginally better than those for the simple relations, aside from the *fu* classifier whose highest score, 0.750, was achieved with the combination of all features for the simple relations. Even though the statistical analysis of the PDTB output at first did not seem promising, the PDTB features did improve all classifiers' performances. The model's performance was best for the majority of classifiers with the features employed by Afantenos et al. (2018) in collaboration with our features for both settings. In particular, our model achieved promising improvements on the attachment and function classifiers.

For illustration, Figure 1 shows the various analyses for one text from the corpus.

6 Discussion and conclusion

In our study we experimented with using discourse parser output for argumentation mining, using presegmented text. We not only looked at RST features, which have already been used in related research, but also experimented with PDTB features. After experimenting with various available parsers, we selected one for RST and one for PDTB, converted their output for our corpus to a common format, and determined correlations. In a follow-up experiment, we used features from both discourse parsers for predicting the argumentation structure, based on a re-implementation of the system of Afantenos et al. (2018). Despite the fact

Feature description	Classifier	Tag
Absolute & relative no. of all children/parents and grandchildren/grandparents of segment	fu, ro	RST+
Relative no. of grandchildren/grandparents before & after the segment	fu, ro	RST+
Absolute & relative distance to parent and direction	at	RST+
Whether the segment is involved in a multi-nuclear relation	at	RST+
Whether segment has any PDTB connections to neighbouring segments	cc, fu, ro, at	PDTB
Count of incoming & outgoing PDTB connectives	cc, fu, ro	PDTB
Level one and two of the PDTB semantic relation	cc, fu, ro, at	PDTB
Raw text of PDTB connective	cc, fu, ro, at	PDTB

Table 4: Feature descriptions.

features	cc	ro	fu	at
Default features	0.722 (+/- 0.068)	0.467 (+/- 0.054)	0.224 (+/- 0.015)	0.673 (+/- 0.034)
Default, RST	0.729 (+/- 0.068)	0.600 (+/- 0.049)	0.278 (+/- 0.034)	0.680 (+/- 0.033)
Default, RST, RST+	0.732 (+/- 0.068)	0.582 (+/- 0.049)	0.305 (+/- 0.048)	0.685 (+/- 0.026)
Default, PDTB	0.771 (+/- 0.073)	0.720 (+/- 0.048)	0.420 (+/- 0.056)	0.691 (+/- 0.030)
Default, RST, RST+, PDTB	0.759 (+/- 0.078)	0.721 (+/- 0.045)	0.417 (+/- 0.050)	0.703 (+/- 0.031)
Default, 2018	**0.854 (+/- 0.057)**	**0.737 (+/- 0.052)**	0.444 (+/- 0.044)	0.720 (+/- 0.023)
Default, 2018, RST, RST+, PDTB	0.852 (+/- 0.054)	0.728 (+/- 0.056)	**0.461 (+/- 0.044)**	**0.732 (+/- 0.027)**

Table 5: Results for the full relation set with complex setting: macro-averaged F1 score, variance in parentheses, maximum is in bold for each classifier

that the PDTB parser only identified a relatively small amount of relations, and these did not map very well to the ARG annotation, the PDTB features still improved the results more than the RST features did (compare lines 2 and 4 to line 1 in Table 5). Combining both feature sets led to further improvements (lines 5, 7). We thus conclude that discourse parser features, and specifically PDTB features, add valuable information in particular for the classification of the function and attachment subtasks of ARG parsing, and could therefore be further explored and applied to other argumentative corpora.

Future work in this line of research includes a qualitative error analysis of the parsers' contributions to ARG parsing, and an ablation test for identifying the impact of the individual RST and PDTB features. Furthermore, recently a second part of the microtext corpus has been released (see website in footnote 1), which is larger than part 1 and would also warrant similar experiments. This would also be a test for the potential influence of the translation step (German to English) in creating part 1.

Acknowledgments

We thank the anonymous reviewers for their helpful comments on the earlier version of the paper.

References

Stergos Afantenos, Andreas Peldszus, and Manfred Stede. 2018. Comparing decoding mechanisms for parsing argumentative structures. *Argument and Computation*, 9(3):177–192.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Or Biran and Kathleen McKeown. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 96–104.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes

and back: Relations and differences. *Lecture Notes in Computer Science (Computational Logic in Multi-Agent Systems)*, 8143:1–17.

Debopam Das and Maite Taboada. 2014. *RST Signalling Corpus Annotation Manual*. Simon Fraser University.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2236–2242.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 511–521.

Nancy Green. 2010. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24(2):181–196.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 13–24.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Ziheng Lin, Hwee Tou Nh, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

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.

Elena Musi, Tariq Alhindi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. 2018. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of the 11h International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan.

Michael O'Donnell. 2000. Rsttool 2.4 a markup tool for rhetorical structure theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253–256.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 938–948.

Andreas Peldszus and Manfred Stede. 2016a. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action - Proceedings of the 1st European Conference on Argumentation 2*, pages 801–816.

Andreas Peldszus and Manfred Stede. 2016b. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Heres my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1364–1373.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–660.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jrmy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 1051–1058.

Douglas Walton, C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.