More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing

Maria Skeppstedt[1,3], **Andreas Peldszus**[2], **Manfred Stede**[3]
[1]Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
maria.skeppstedt@lnu.se
[2]Retresco GmbH, Berlin, Germany
andreas.peldszus@retresco.de
[3]Applied Computational Linguistics, University of Potsdam, Potsdam, Germany
stede@uni-potsdam.de

Abstract

We present an extension of an annotated corpus of short argumentative texts that had originally been built in a controlled text production experiment. Our extension more than doubles the size of the corpus by means of crowdsourcing. We report on the setup of this experiment and on the consequences that crowdsourcing had for assembling the data, and in particular for annotation. We labeled the argumentative structure by marking claims, premises, and relations between them, following the scheme used in the original corpus, but had to make a few modifications in response to interesting phenomena in the data. Finally, we report on an experiment with the automatic prediction of this argumentation structure: We first replicated the approach of an earlier study on the original corpus, and compare the performance to various settings involving the extension.

1 Introduction

As with most areas in NLP, progress on Argumentation Mining hinges on the availability of data, and in the case of this field, this is generally taken to be *annotated* data. Up to now, only few corpora labelled with full argumentation structure (i.e., argument components and relations between them) are available; prominent ones are the persuasive essay corpus of Stab and Gurevych (2014), the web text corpus of Habernal and Gurevych (2017), and the argumentative microtext corpus of Peldszus and Stede (2016).[1] The latter is interesting because it has in parallel been annotated with various other linguistic layers, as will be described in Section 2. The microtexts are relatively "clean" text, and the annotation of argumentation structure was generally easy, leading to reasonable annotator agreement, as reported by Peldszus and Stede (2016). However, a drawback is the relatively small corpus size: 112 texts of about five argumentative text units on average. While this data has proven to be useful for various purposes (see Section 2), for machine learning it is clearly desirable to have a larger corpus of this kind.

In this paper, we turn to crowdsourcing as a means to generate more text. We used essentially the same instructions as used by Peldszus and Stede (2016), and recruited writers via Amazon Mechanical Turk. Naturally, the set of resulting texts is not identical in nature to the original ones, and thus the first contribution of this paper is an analysis of how the different text elicitation scenarios influences the outcome, i.e., to evaluate the pros and cons of crowdsourcing for this type of task. The second contribution is an evaluation of the annotation scheme that was used for argumentation: Which modifications are necessary in order to produce adequate analyses of the text? Finally, the third contribution is to report on results of an automatic classification experiment: We replicated the Minimum Spanning Tree approach proposed by Afantenos et al. (2018), and we compare the results that have already been achieved on the original corpus to those stemming from the new sections of the corpus. We regard this as valuable information on the influence of corpus size on classification results.

In the following, as background we briefly describe the original corpus, and then explain our approach to crowdsourcing the text production task. This is followed by a description of the annotation phase, and the lessons learned. Finally, we report on the classification experiment, and then sum up.

The new corpus data, with its annotation of argumentation structure, is available on the website of the arg-microtext corpus (see below).

[1]Many other corpora are available with more lean or more specific annotations; see Section 4 of (Lippi and Torroni, 2016).

Proceedings of the 5th Workshop on Argument Mining, pages 155–163
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

2 Background: The 'argumentative microtext corpus'

2.1 Data

We start from the arg-microtext corpus (Peldszus and Stede, 2016), a freely available[2] parallel corpus of 112 short texts with 576 argumentative discourse units (henceforth: segments). The texts are authentic discussions of controversial issues, which were given to the writers as prompts. They were originally written in German and have been professionally translated to English, preserving the segmentation and if possible the usage of discourse markers. The texts have been collected in a controlled text generation experiment, in a classroom setting with students, using a short instruction. This had the result that all of the texts fulfill the following criteria: (i) The length of each text is about 5 segments; (ii) one segment explicitly states the central claim; (iii) each segment is argumentatively relevant; (iv) at least one objection to the central claim is considered (in order to produce more interesting argumentation).

Finally, all texts have been checked for spelling and grammatical problems, which have been corrected by the annotators. The reason underlying this decision was the intended role of the corpus as a resource for studying argumentation in connection with other linguistic phenomena (see Section 2.3), where plain errors can lead to undesired complications for parsers, etc. Hence, "authenticity" on this level was considered as less important. In this respect the corpus differs from web-text corpora that have been collected for argumentation mining purposes, such as the Internet Argument Corpus (Abbott et al., 2016), the ABCD corpus (Rosenthal and McKeown, 2015) and others.

2.2 Annotation scheme

The argumentation structure of every text was annotated according to a scheme proposed by Peldszus and Stede (2013), which in turn had been based on Freeman's theory of argumentation structures (Freeman, 2011). This annotation scheme has already been proven to yield reliable structures in annotation and classification experiments, for instance by (Peldszus and Stede, 2015; Potash et al., 2017). (Stab and Gurevych, 2017) use a similar scheme for their corpus of persuasive essay, and they also provide classification results for the microtext corpus.

The argumentation structure of a text is defined as a tree with the text segments as nodes. Each node is associated with one argumentative role: the *proponent* who presents and defends the central claim, or the imaginary *opponent* who critically questions the proponent's claims. Edges between the nodes represent argumentative relations: *support* or *attack*. The scheme allows to discriminate between 'rebutting' attacks, targeting another node and thereby challenging its acceptability, and 'undercutting' attacks, targeting an edge and thereby challenging the acceptability of the inference from the source to the target node. It can also represent linked support, where multiple premises jointly support a claim, i.e., one of the premises would not be able to play the support role in isolation. Another category is 'example support', where the supporting material is a concrete instance of some abstract proposition, serving as evidence. Finally, it is possible to identify two segments as saying essentially the same thing, hence the second being a restatement of the first. (This typically occurs with central claims, which are sometimes being rephrased at the end of the text.)

For illustration, sample analyses are shown below in Figures 1 and 2.

2.3 Other annotation layers

In contrast to other argumentation corpora, the microtext corpus is unique in that it is already annotated with further layers of linguistic information, which makes it usable for systematic correlation studies. Stede et al. (2016) described the annotation of discourse structure according to RST and SDRT, and Becker et al. (2016) added information on *situation entity types*, which Smith (2003) had proposed as a linguistic tool for identifying different 'discourse modes', viz. Narrative, Description, Report, Information, and Argument. Reisert et al. (2017) annotated part of the corpus with information on argumentation schemes, in the spirit of Walton et al. (2008). Also, an alternative approach to schemes, that of Rigotti and Greco Morasso (2010), was annotated on the microtexts by Musi et al. (2018).

Given these extra layers, we regard the extension of the microtext corpus as especially useful, as the annotations of the other layers may now also be added, resulting in a much more valuable re-

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

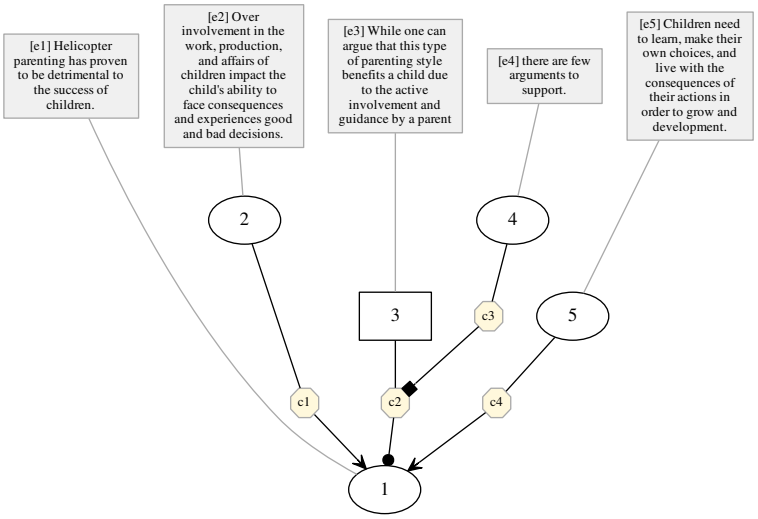

Figure 1: An example text and its argumentation structure: Text segments, proponent (round) and opponent (box) nodes, supporting (arrow-head) and attacking (circle-head) relations.

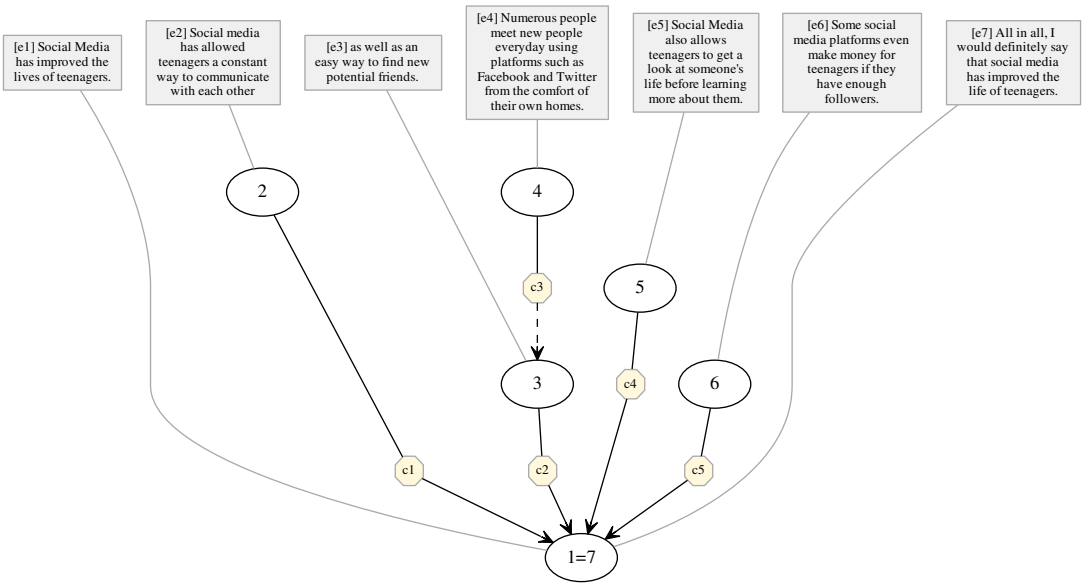

Figure 2: An example of an argumentation structure for which the main claim is repeated in the text. Each segment has been annotated as an independent argumentative discourse unit, all of them directly supporting the main claim, with the exception of one unit which gives support by example.

source, both in terms of volume and in terms of phenomena that can be investigated.

3 Crowdsourcing the production of argumentative texts

3.1 Setting

We recruited authors via Amazon Mechanical Turk, making sure (as far as possible) that they were High School Graduates and living in the U.S. (to increase the chances for language competence). The authors were given the task of producing a short text that argued for or against a general debate topic, the prompt. Everybody was given one from the set of 36 prompts and wrote no more than one text for the experiment. Prompts were gathered from publicly available essay-writing exercises, making sure that they do not presuppose local or temporally-restricted knowledge that our Turkers might not have. For illustration, here are three of the prompts we used:

- Should car drivers be strictly prohibited from using their cell phones?
- Does recycling really make a difference?
- Do older people make good or bad parents?

We calculated the time it takes authors on average by means of a pilot study, and then decided to pay the authors 1.10$ for their effort. Like in the original setting descibed in the previous section, authors were instructed to use about 5 sentences, to clearly take a stance and make it explicit by means of a claim statement, and to also include at least one argument for the opposite view in their text.

3.2 Filtering

As to be expected, not all of the texts that were produced did, however, fulfill all the criteria. First, a number of them did not mention the opposing view; since this does not lead to degenerate data in any way, we decided to leave those texts in the corpus. In contrast, texts in which no clear stance towards the debate topic was taken were excluded from further annotation. Such texts typically listed a number of conditions for agreeing with the topic at hand, gave recommendations for solving an issue, or simply listed a few arguments for and against the topic, without indicating a winning side. Also, we removed texts where authors took a stance, but mainly wrote things unrelated to the debate. Likewise, texts that were not understandable (for grammatical and/or content reasons), and texts that were very long or very short (more than eight or less than four argumentative discourse units) were excluded. Since the debate topics given were very general, the authors sometimes voiced a more specific opinion on a topic. For instance, for the prompt "Do long distance relationships work?", two of the authors argued for the more specific stance: "long-distance relationships work in the short run but not in the long run". These texts, being otherwise faithful to our criteria, were kept in the corpus for annotation.

3.3 Cleaning

The texts that we kept after the filtering phase were manually cleaned, i.e., minor misspellings and grammatical errors were corrected. Furthermore, some of the authors have taken a clear stance on the given prompt question by simply starting their text with "yes" or "no", before presenting their arguments. This violates the guidelines, which ask for texts that should be understandable without actually having the question as headline. For these cases, the answer was replaced with a statement that paraphrased the prompt question and indicated the "yes" or "no" polarity. In addition, text-initial anaphors (referring to parts of the question) were replaced with their intended antecedents.

We are aware that cleaning and repairing are potentially controversial moves. Our main motivation was that the data be comparable to that of the original corpus, and therefore we largely followed the 'cleaning' procedure described by Peldszus and Stede (2016). However, all "raw" versions of the texts will also be part of the corpus release, as for certain experiments it might be important to be confronted with authentic language containing mistakes of various kinds.

3.4 Statistics

A total of 205 texts had been originally collected, and from these, 34 were excluded from further consideration, for the reasons given above (but still part of the corpus to be distributed). Thus, we altogether moved 83% of the crowdsourced texts to the next phase: annotation of argumentation structure. We see this rate as rather encouraging, demonstrating that crowdsourcing is a viable approach for this type of text elicitation task.

4 Annotating the crowdsourced texts

We applied the annotation guidelines (mentioned in Section 2) and used the freely available graph annotation tool GraPAT[3] (Sonntag and Stede, 2014) to annotate the 171 texts that passed our filtering step. Two annotators (one of them being a co-author of this paper) shared the work, and a third person (another co-author) joined in discussions of difficult cases. At the present stage, we did not run an inter-annotator agreement study, because this had already been done on the original corpus and guidelines (see Peldszus and Stede (2016)), thereby verifying the usability of the scheme. However, the annotation process was not entirely straightforward. In the following, we describe specifically the challenges posed by the different type of text, in comparison to the original microtexts and the annotation scheme. We regard most of the phenomena as not just specific to this project, but to be relevant for empirical work on argumentation mining, especially for designing annotation guidelines, in general.

4.1 Implicit claims

The first observation concerns the presence of an explicit "central claim". Authors were encouraged to state it in their text, but we did not filter texts that lack it (because, in fact, many "natural" argumentative texts have no explicit claim, as for instance found by Habernal and Gurevych (2017)). As long as the argumentative structure and content of the text suggested some segment to be a viable candidate for main claim, our annotators chose it. This had the effect that—in contrast to the original corpus with its rather crisp claims—both specific refinements of the writing prompt (e.g., "As long as the kids are provided with a stable home life, divorce does not have to be an enormous trauma from which there is no recovering") and relatively vague statements (e.g., "There are many benefits to using LED lights") can now be central claims. However, when the text argued clearly, but did not supply any reasonable candidate for explicit claim, annotators added this as an extra statement, which will then serve as the root of the argumentation tree. A manually added statement serving as the main claim was added in 34 texts. One example is the following text, where the last sentence is the manually added claim:

(1) Do we need fracking, depite its risks? Fracking has uncovered cheap natural gas. The aggregate savings to the American household are then passed on to the economy in the way of spending. Also, the coal industry has imploded as a consequence which is more of a pollutant than natural gas. The potential contamination damage caused by the fracking process is outweighed by the reduction of energy costs to the American household. Yes, we need fracking despite its risks.

4.2 Restatements

Another phenomenon that did not occur in the previously published corpus, was that the authors restated a claim, typically the main claim. These restatements were annotated through connecting the text segments that restated the claim to the same argumentative discourse unit, as shown in Figure 2 (node '1 = 7'). In the annotated corpus, 29 argumentative units are restatements of previously mentioned ones, and 19 of them restate the main claim.

4.3 Direct versus indirect support

Another difficulty concerned the attachment point of support relations. It can be difficult to decide whether a statement supports or opposes directly the central claim, or a separate statement (thus affecting the claim only indirectly). This kind of ambiguity was also reported for the arg-microtext corpus by Peldszus and Stede (2016). We noted that it appears quite frequently in the crowdsourced texts. For instance, in Figure 2, all segments, except one, are annotated as direct support of the main claim, as there are no surface markers (or clear semantic cues involving background knowledge) in the text which would signal that any of these arguments support any other argument. However, the author may have intended additional supporting relations.

4.4 Argument support versus causal connection

Another challenge stems from drawing the line between relations in the texts that are argumentative, and those that describe a (non-pragmatic) causal connection of events. The example below may be viewed as one single argumentative discourse unit, which includes one long causal connection. Alternatively, it may be segmented into

[3] http://angcl.ling.uni-potsdam.de/resources/grapat.html

three sub-arguments, on the ground that it is possible to agree on or refute each one of these three segments separately. E.g., it is possible to agree that people go to the stores to recycle, but refute that this leads to more money being spent in the shop, or that this leads to economic growth.

(2) It is also a benefit as it encourages people to go to stores to recycle and then spend that money at the shop increasing the amount spent at the store and encouraging economic growth.

4.5 Implicit modality or evaluation

In many cases, annotation decisions turned out to be dependent on whether a certain modality or a positive or negative evaluation is added to a segment by the annotator's interpretation. In the (partial) text below, segment 2, 3 and 4 are annotated as supporting the claim 1. In turn, 5 supports 4, given that the annotator interprets "a heart rate that gets going" as a positive state of affairs, brought about by the desire to keep weight.

(3) [Spending time together as a family engaged in sports together is a good thing.]$_1$ [It increases a sense of family togetherness,]$_2$[gets people outside and into the fresh air and sunshine,]$_3$[and gets the heart rate going.]$_4$[This in turn helps to keep weight at a healthy level]$_5$ (...)

4.6 Non-argumentative text units

Finally, there were three cases in which the texts contained segments that the annotators deemed to be irrelevant for the argument, for instance because it provides only background information or reports some personal experience of the author that is only vaguely related. In the original corpus this was not the case and hence lead to tree structures spanning the entire text. Now in the crowdsourced texts, we decided to leave those texts in the corpus and therefore now have segments that are not part of the graph. An example is the beginning of a text on the pros and cons of soft drink can deposits:

(4) I live in Michigan, where we have a deposit.

4.7 Summary

The annotation effort lead to a total of 932 argumentative units (segments). The distribution of relations is: convergent support (467); example support (23); rebutting attack (137); undercutting attack (77); linked support or attack (57); restatement (29).

5 Experiments on automatic classification

In the following, we will describe our experiments on automatically identifying the argumentative structures. This has already been done on the original version of the corpus, e.g., recently by Afantenos et al. (2018). In our experiments we replicate their approach, and test it on the texts we acquired and annotated as described above. Our aim is to get an understanding of how much the old and the new data sets differ in terms of achievable predictions, and to assess possible improvements by extending the size of the corpus.

Regarding the new phenomena pointed out in Section 4, we chose to ignore non-argumentative segments for the purposes of this experiment, similar as if they had been filtered out in a prior step of a pipeline. For one thing, this concerns only three texts, and, more importantly, if we want to compare our results to the earlier work, we should work with the same representations. Second, implicit claims that have been made explicit by the annotators are included when we predict argumentation structures.

5.1 Experimental Setup

For predicting argumentation structures, we replicated the MST model of Afantenos et al. (2018), which is an improved version of the model originally presented by Peldszus and Stede (2015). This approach learns four local models for various aspects of the argumentation structure: for identifying the central claim of the text (cc); for determining the argumentative role, i.e. proponent or opponent (ro), for classifying the function of a segment, such as support or attack (fu); and finally for identifying which units are 'attached' to each other, i.e. are connected by an argumentative relation (at). The predictions of these local models are then combined into a single edge score and decoded to a structure by selecting the minimum spanning tree (MST). This approach has been shown to yield competitive results when compared to ILP decoders; see the original papers for more details.

Similar to previous work, our experiment uses the argumentation graphs in a version that is con-

verted to dependency structures. Also, the set of relations is reduced to merely 'support' and 'attack' by conflating the subcategories. This step is done in order to be compatible with earlier work (no other corpora use this set of fine-grained distinctions of support and attack so far) and to alleviate a potential sparse-data problem per specific relation. Restatements (which tend to occur only for the main claim) exist in the new data set but not in the original one; for compatibility, we converted them to support relations in order to maintain compatibility with the old corpus. Again, this is a purely technical decision made in order to allow a comparison with prior and related work. As an alternative, experiments with the fine-grained set of relation have been done (on the original corpus) by Peldszus (2018).

We adopt the evaluation procedure of previous work, i.e., we use 50 train-test splits, resulting from 10 randomized repetitions of 5-fold cross validation. For evaluations on the original corpus we use the published splits, for the new corpus we derive splits analogously. The correctness of predicted structures is measured separately for the four subtasks, reported as macro averaged F1, and more unified in a labelled attachment score (LAS) as it is commonly used for evaluation in dependency parsing (see Kübler et al., 2009, ch. 6.1). For significance testing, we use the Wilcoxon signed-rank test (Wilcoxon, 1945).

5.2 Evaluation Scenarios

We compare the results on the original dataset and those on the new one using three evaluation scenarios:

Single Corpus This is the standard scenario for evaluating the model on one single corpus, from which both training and test sets are sampled. We reproduce the results on the original corpus, and produce new results for the new corpus. Comparing these scores gives a first, but only tentative, impression whether the structures annotated in the new corpus are as easy or as hard to recognize as in the original corpus.

Cross Corpus When we train the model exclusively on one corpus and test it on the other, we can investigate the degree of generalization of the model. This is especially interesting, since the new corpus had different prompts and thus covers different topics. We expect a decrease in performance when compared to in-domain results as in the single-corpus setting.

Extend Corpus Finally, we use one corpus as additional training data when evaluating on the other. This helps us to understand to which degree new data can help achieve better results for the four subtasks and overall for the prediction of full structures. We expect improvements here, when compared with the single-corpus setting.

5.3 Results

The results are shown in Table 1. In the scenario, 'old' refers to the original corpus, and 'new' to the new one described in this paper.

Single Corpus The results reproduced on the original corpus are equivalent to published results. On the new corpus, we overall achieve similar scores. Differences are subtle: central claims are a bit harder to recognize (an absolute difference of -2.5 points) on the new corpus. This is to be expected, as the new corpus features restatements of the main claim which are competitors to the original main claim. The scores for argumentative role, function and attachment classification are quite equal. This leads us to assume that the structures annotated in the new corpus are not more or less complicated to be recognized than the structures in the original corpus.

Cross Corpus As expected, the cross-corpus results are in general lower than single-corpus scores for both directions. When training on the old corpus and testing on the new one, we observe a relative decrease of 7% compared to the average level score achieved when training and testing on the new corpus. The loss is slightly stronger for argumentative function and attachment than on the other levels. In the reverse direction, when training on new and testing on old, the average loss is even higher with 11%. Here, central claim and argumentative function exhibit the highest decrease. The exception is the attachment level, with only a minor drop of 3%.

Extend Corpus When using the "other" corpus as additional training data and comparing this with the 'single' scenario without extra training data, we find on average only mild improvements (which we again report as relative improvements). Interestingly, the gains per task differ across the directions: When evaluating on the old corpus using the new data for extra training, there is a small

scenario			results				
type	train	test	cc	ro	fu	at	LAS
single	old	old	.870	.768	.754	.719	.526
cross	new	old	.745	.695	.644	.698	.450
extend	both	old	.859	.779†	.757	.724	.532
single	new	new	.845	.766	.750	.714	.527
cross	old	new	.797	.731	.693	.665	.439
extend	both	new	.856†	.782†	.765‡	.712	.526

Table 1: Evaluation scores for the predicted structures reported as macro avg. F1 for the cc, ro, fu, and at levels, and as labelled attachment score (LAS). Results marked with a dagger are significant improvements over the corresponding 'single' score, with † for $p < 0.05$ and ‡ for $p < 0.01$.

drop (-1.3%) in central claim identification and a small raise in role classification (+1.4%). The remaining levels show minor improvements. In the other direction, i.e. when evaluating on the new corpus using the old corpus as additional training data, we observe improvements in role (+2.1%) and function classification (+2.0%), as well as a small raise in central claim identification (+1.3%). One possible explanation for this is the impact of the restatements in the new corpus. An improvement that is consistent across both directions is that in role classification. We presume that more training data really helped to recognize the less frequent opponent role.

6 Summary and Outlook

In order to extend an existing corpus of 112 short argumentative texts (which had been gathered in a classroom setting with students), we employed crowdsourcing for collecting a new dataset that can serve as an extension to the old one. We described our steps in assembling the data set in such a way that is compatible to the original corpus but at the same time is to some extent faithful to the "crowdsource complications". As a result, there are two changes in the corpus now: Texts may contain non-argmentative segments, and some "artificial" segments representing central claims have been added where authors left the claim implicit. Still, these are no dramatic steps, and overall, we claim that (i) crowdsourcing can be a viable method for collecting this type of data, and that (ii) the new corpus can be used in tandem with the old one as a coherent dataset.

Finally, to substantiate (ii), we reproduced an experiment on automatic prediction of the argumentation structure, which showed that predicting on the crowdsourced texts is generally not harder than on the old ones, and that overall, the task can benefit from the increased corpus size, though not dramatically. But we expect the increased corpus size to be useful for other machine learning experiments, especially for neural network approaches, such as those recently run by Potash et al. (2017) on the old corpus (albeit using only a small part of the annotations for a simplified setting).

An interesting question for future work concerns the viability of using crowdsourcing not just for collecting the texts, but also for annotation. Instead of having annotators draw graph structures, one would translate the process into a sequence of questions whose answers would imply the structural description. We plan to explore this path with suitable pilot experiments.

The corpus and annotations are available from the arg-microtexts website (see footnote 2 above).

Acknowledgements

We would like to thank Constanze Schmitt for carrying out annotations and for translating the annotation guidelines into English, and Anna Laurinavichyute for setting up and running the MTurk experiment. We would also like to thank the Swedish Research Council (Vetenskapsrådet) that partly funded this study through the project "Navigating in streams of opinions: Extracting and visualising arguments in opinionated texts" (No. 2016-06681). Finally, thanks to the anonymous reviewers for their constructive comments.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proc. Language Resources and Evaluation*, pages 4445–4452.

Stergos Afantenos, Andreas Peldszus, and Manfred Stede. 2018. Comparing decoding mechanisms for parsing argumentative structures. *Argument and Computation*. (published online, February 2018).

Maria Becker, Alexis Palmer, and Anette Frank. 2016. Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argumentation Mining*, Berlin. Association for Computational Linguistics.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Argumentation Library (18). Springer, Berlin/New York.

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

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency Parsing. *Synthesis Lectures on Human Language Technologies*, 2(1):1–127.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

Elena Musi, Smaranda Muresan, and Manfred Stede. 2018. A multilayer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Andreas Peldszus. 2018. *Automatic recognition of argumentation structure in short monological texts*. Ph.D. thesis, Universität Potsdam.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 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*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–816, London. College Publications.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1375–1384, Copenhagen, Denmark. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Deep argumentative structure analysis as an explanation to argumentative relations. In *Proceedings of The 23rd Annual Meeting of the Association for Natural Language Processing*, pages 38–41.

Eddo Rigotti and Sara Greco Morasso. 2010. Comparing the argumentum model of topics to other contemporary approaches to argument schemes: The procedural and material components. *Argumentation*, 24(4):489–512.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.

Carlota Smith. 2003. *Modes of discourse. The local structure of texts*. Cambridge University Press, Cambridge.

Jonathan Sonntag and Manfred Stede. 2014. Grapat: a tool for graph annotations. In *Proceedings of LREC 2014*, Reykjavik.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

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

D. Walton, C. Reed, and F. Macagno. 2008. *Argumentation schemes*. Cambridge University Press, Cambridge.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.