Unshared Task at the 3rd Workshop on Argument Mining: Perspective Based Local Agreement and Disagreement in Online Debate

Chantal van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo and **Piek Vossen**
Vrije Universiteit Amsterdam
Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{c.m.van.son, t.caselli, antske.fokkens, isa.maks, r.morantevallejo, lora.aroyo, piek.vossen}@vu.nl

Abstract

This paper proposes a new task in argument mining in online debates. The task includes three annotations steps that result in fine-grained annotations of agreement and disagreement at a propositional level. We report on the results of a pilot annotation task on identifying sentences that are directly addressed in the comment.

1 Introduction

Online debate (in its broadest sense) takes an increasingly prominent place in current society. It is at the same time a reflection and a shaping factor of the different beliefs, opinions and perspectives that exist in a certain community. Online debate characterizes itself by the dynamic interaction between its participants: they attack or support each other's stances by confirming or disputing their statements and arguments, questioning their relevance to the debate or introducing new arguments that are believed to overrule them. In fact, as Peldszus and Stede (2013, p. 4) point out, all argumentative text is of dialectic nature: "an argument always refers to an explicitly mentioned or at least supposed opponent, as for instance in the rebutting of possible objections." Therefore, these (implicit) interactions between participants should be given a central role when performing argument mining.

In recent years, several studies have addressed the annotation and automatic classification of *agreement* and *disagreement* in online debates. The main difference between them is the annotation unit they have targeted, i.e. the textual units that are in (dis)agreement. Some studies focused on *global* (dis)agreement, i.e. the overall stance towards the main debate topic (Somasundaran and Wiebe, 2010). Other studies focused on *local* (dis)agreement, comparing pairs of posts (Walker et al., 2012), segments (Wang and Cardie, 2014) or sentences (Andreas et al., 2012). Yin et al. (2012) propose a framework that unifies local and global (dis)agreement classification.

This paper describes an argument mining task for the Unshared Task of the 2016 ACL Workshop on Argument Mining,[1] where participants propose a task with a corresponding annotation model (scheme) and conduct an annotation experiment given a corpus of various argumentative raw texts. Our task focuses on local (dis)agreement. In contrast to previous approaches, we propose *micro-propositions* as annotation targets, which are defined as the smallest meaningful statements embedded in larger expressions. As such, the annotations are not only more informative on exactly *what* is (dis)agreed upon, but they also account for the fact that two texts (or even two sentences) can contain both agreement and disagreement on different statements. The micro-propositions that we use as a basis have the advantage that they are simple statements that can easily be compared across texts, whereas overall propositions can be very complex. On the other hand, creating a gold-standard annotations of micro-propositions is time consuming for long texts. We therefore propose an (optional) additional annotation step which identifies relevant portions of text. This results in a three-step annotation procedure: 1) identifying relevant text, 2) identifying micro-propositions and 3) detecting disagreement. We report on a pilot study for the first subtask.

We selected a combination of two data sets provided by the organizers: i) Editorial articles extracted from Room for Debate from the N.Y. Times website (Variant C), each of which has a debate title (e.g. *Birth Control on Demand*), debate description (e.g. *Should it be provided by the gov-*

[1] http://argmining2016.arg.tech/index.php/home/call-for-papers/unshared-task

Proceedings of the 3rd Workshop on Argument Mining, pages 160–165,
Berlin, Germany, August 7-12, 2016.

ernment to reduce teen pregnancies?) and article title describing the author's stance (e.g. *Publicly Funded Birth Control Is Crucial*); and ii) Discussions (i.e. collections of comments from different users) about these editorial articles (Variant D).

The remainder of this paper is structured as follows. Section 2 introduces the theoretical framework the task is based on. The annotation task is described in Section 3. Section 4 discusses the results of an annotation experiment, and we conclude and present future work in Section 5.

2 Perspective Framework

We consider any (argumentative) text to be a collection of propositions (statements) associated with some *perspective values*. In our framework (van Son et al., 2016), a *perspective* is described as a relation between the source of a statement (i.e. the author or, in the case of quotations, another entity introduced in the text) and a target in that statement (i.e. an entity, event or proposition) that is characterized by means of multiple perspective values expressing the attitude of the source towards the target. For instance, the commitment of a source towards the factual status of a targeted event or proposition is represented by a combination of three perspective values expressing *polarity* (AFFIRMATIVE or NEGATIVE), *certainty* (CERTAIN, PROBABLE, POSSIBLE) and *time* (FUTURE, NON-FUTURE). Other perspective dimensions, such as sentiment, are modeled in the same way with different sets of values.

Our assumption is that participants in an online debate interact with each other by attacking or supporting the perspective values of any of the propositions in a previous text. In this framework, we define *agreement* as a correspondence between one or more perspective values of a proposition attributed to one source and those attributed to another source; *disagreement*, on the other hand, is defined as a divergence between them. For example, consider the following pair of segments, one from an editorial article and the other from a comment in the context of *Teens Hooked on Screens*:

> **ARTICLE:** The bullies have moved from the playground to the mobile screen, and there is no escaping harassment that essentially lives in your pocket.
>
> **COMMENT:** Ms. Tynes: The bullies haven't moved from the playground to the screen.

This is a clear example of disagreement on the perspective values of a proposition present both in the editorial article and in the comment. As represented in Figure 1, the article's author commits to the factual status of the proposition, whereas the commenter denies it. In this example, the disagreement concerns the whole proposition ("no moving took place at all"). However, we assume that (dis)agreement can also target specific arguments within a proposition (i.e. hypothetically, someone could argue that it is not the bullies that moved from the playground to the screen, but someone else). We call these smallest meaningful propositional units in a text *micro-propositions*.

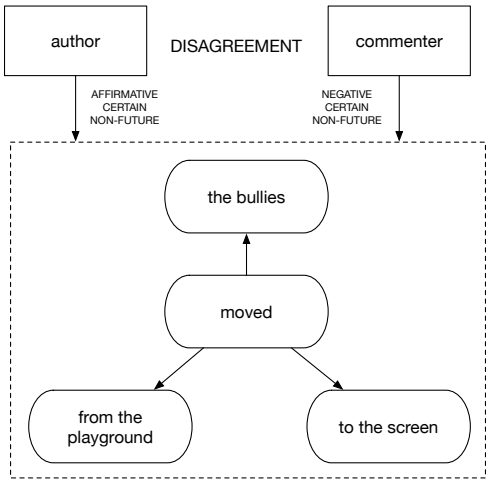

Figure 1: Representation of disagreement in the perspective framework.

3 Task Definition

Based on the perspective framework, we propose a task that aims at determining whether authors agree or disagree on the perspective values associated with the propositions contained in debate texts. Rather than trying to model the full debate comprehensively, we propose to start from the smallest statements made (i.e. micro-propositions) and derive more overall positions from the perspectives on these statements. This requires a detailed analysis of the texts. We optimize the annotation process by dividing the task into three subtasks described below: (1) Related Sentence Identification, (2) Proposition Identification, and (3) Agreement Classification.

3.1 Task 1: Related Sentence Identification

In an online debate, people often do not respond to each and every statement made in previous texts, but instead tend to support or attack only one or a few of them. The aim of the first task is to identify those sentences in the editorial article that are

COMMENTED_UPON in the comment. A sentence is defined to be COMMENTED_UPON if:

- the comment **repeats** or **rephrases** (part of) a statement made in the sentence;
- the comment **attacks** or **supports** (part of) a statement made in the sentence.

The main purpose of this task is to eliminate the parts of the editorial article that are irrelevant for (dis)agreement annotation. In the data set we use in this paper, the average number of sentences in the editorial articles is 19 (in the comments, the number of sentences ranges from 1 to 16). Without this first task, all propositions of the article including the irrelevant ones would have to be identified and annotated for (dis)agreement, which is neither efficient nor beneficial for the attention span of the annotators. With other data consisting of short texts, however, this subtask may be skipped.

Deciding whether a statement is COMMENTED_UPON may require some reasoning, which makes the task inherently subjective. Instead of developing overdetailed annotation guidelines simply to improve inter-annotator agreement, we adopt the view of Aroyo and Welty (2014) that annotator disagreement can be an indicator for language ambiguity and semantic similarity of target annotations. We considered using crowdsourcing for this task, which is particularly useful when harnessing disagreement to gain insight into the data and task. However, platforms like CrowdFlower and MTurk are not suitable for annotation of long texts and eliminating context was not an option in our view. Therefore, the task is currently designed to be performed by a team of expert annotators, and we will experiment with different thresholds to decide which annotations should be preserved for Tasks 2 and 3. In the future, we might experiment with alternative crowdsourcing platforms.

3.2 Task 2: Proposition Identification

A sentence can contain many propositions. For instance, the article sentence discussed earlier in this paper (repeated below) contains three propositions centered around the predicates marked in bold:[2]

> **ARTICLE:** The bullies have **moved** from the playground to the mobile screen, and there is no **escaping** harassment that essentially **lives** in your pocket.

[2]We do not consider *is* to express a meaningful proposition in this sentence.

In Task 2 we annotate the (micro-)propositions in the sentences that have been annotated as being COMMENTED_UPON. We first identify predicates that form the core of the proposition (e.g. *moved*, *escaping* and *lives*). Next, we relate them to their arguments and adjuncts. For the first predicate *moved*, for example, we obtain the following micro-propositions:

- moving
- the bullies moved
- moved from the playground
- moved to the screen

In this task, we annotate linguistic units. Though we will experiment with obtaining crowd annotations for this task, we may need expert annotators for creating the gold standard. We expect to be able to identify micro-propositions automatically with high accuracy.

3.3 Task 3: Agreement Classification

The final goal of the task is to identify the specific micro-propositions in the editorial article that are commented upon in a certain comment, and to determine whether the commenter agrees or disagrees with the author of the article on the perspective values of these micro-propositions. Thus, the final step concerns classifying the relation between the comment and the micro-propositions in terms of agreement and disagreement. For example, there is disagreement between the author and the commenter about the factual status of *moving*. We aim to obtain this information by asking the crowd to compare micro-propositions in the original text to those in the comment.

Even though most irrelevant micro-propositions have been eliminated in Task 1, we need an IRRELEVANT tag to mark any remaining micro-propositions for which (dis)agreement cannot be determined (e.g. all those obtained for *escaping* and *lives* in our example).

3.4 Interaction between Subtasks

The first two subtasks are primarily used to provide the necessary input for the third subtask. The relation between the second and third task is clear. In the second task, we create the units of comparison and in the third task we annotate the actual (dis)agreement. Similarly, the first task directly provides the input for the second task. The relation between the first and third task is more complex. In order to establish whether a comment

comments upon a specific sentence, we need to determine if there is any (dis)agreement with the sentence in question. A natural question may be how this can be done if this information is only made explicit in subtask 3 or why we need to carry out subtasks 2 and 3 if we already established (dis)agreement in subtask 1. The main difference lies in the level of specificity of the two tasks. In subtask 1, annotators are asked if a comment addresses a given sentence in any way. Subtask 3 dives deeper into the interpretation by asking for each micro-proposition in the sentence whether the commenter agrees or disagrees with it.

There may be cases where one of the subtasks assumes that there is (dis)agreement and the other that there is no relation. We use the following strategies to deal with this. When no (dis)agreement is found on a detailed level, subtask 3 provides an option to indicate that there is no relation between a micro-proposition and the comment (the IRRELEVANT tag). This captures cases that were wrongly annotated in subtask 1. If subtask 1 misses a case of (dis)agreement, this cannot be corrected in subtask 3. We can, however, maximize recall in the first subtask by using multiple annotators and a low threshold for selecting sentences (e.g. requiring only one annotator to indicate whether the sentence is commented upon). We will elaborate on this in Section 4.

4 Task 1: Pilot annotation

This section reports on a pilot annotation experiment targeted at the first subtask. Five expert annotators were asked to identify those sentences in the editorial article that were COMMENTED_UPON in the comment. A set of eight editorial articles (152 unique sentences, including titles) and a total of 62 comments were provided. In total, this came down to 1,186 sentences to be annotated. We used the Content Annotation Tool (CAT) (Lenzi et al., 2012) for the annotations.

The experiment was performed in two rounds. First, simple instructions were given to the annotators to explore the data and task. For the second round, the instructions were refined by adding two simple rules: exclude titles (they are part of the meta-data), and include cases where a proposition is simply 'mentioned' rather than functioning as part of the argumentation. For example, the fact that the closing of Sweet Briar College is repeated in the comment below without its factual status being questioned most likely means that there is an agreement about it, so we do need to annotate it:

> **ARTICLE:** Despite a beautiful campus, dedicated faculty, loyal alumnae and a significant endowment, Sweet Briar College is closing after 114 years.
>
> **COMMENT:** Anyway, there's something ineffably sad to me about Sweet Briar's closing.

Figure 2 shows the distribution of the annotations in both rounds. Only the sentences that were annotated by at least one annotator (29% in Round 2) are included in the graph. We explained in Section 3.1 that identifying whether a sentence is commented upon or not is an inherently subjective task. We analyze the distribution of annotations, because distributions are more insightful for tasks where disagreement is expected than measurements for inter-annotator agreement.

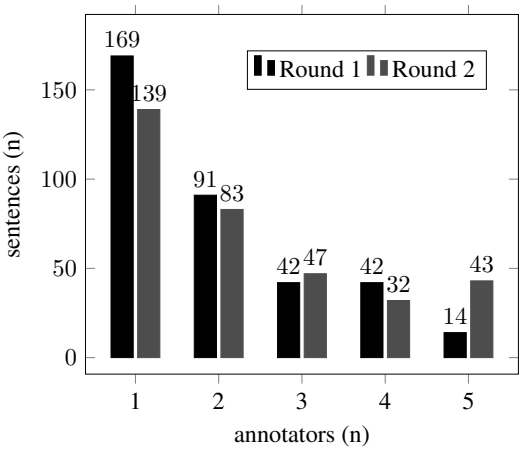

Figure 2: Distribution of annotations.

A deeper analysis of the annotated data and the annotation distribution shows the different degrees of connectivity between the annotated sentences and the comments. The sentences that were annotated by 4 or 5 annotators clearly were strongly and unambiguously related to the comment. For example, the following sentence was annotated by all of the annotators:

> **ARTICLE:** But allowing children and teens to regulate their behavior like adults gives them room to naturally modify their own habits.
>
> **COMMENT:** I empathize with your argument that allows children to regulate their behavior like adults.

In addition, the above sentence is one example of those that were annotated as being COMMENTED_UPON in multiple comments. What these sentences seem to have in common is that they express an important argument or a concluding statement in the editorial article. In the above case, the author of the article uses this argument in

an online debate about *Teens Hooked on Screens* to argue why you should not limit your teen's screen time. Comparing this example to one where only a minority of the annotators agreed (i.e. 2 out of 5), a difference can be noticed in the amount of inference that is required to understand a relation between the sentence and the comment (i.e. the article sentence specifies *how* access to birth control is a win-win for young women):

> **ARTICLE:** Giving poor young women easy access to birth control is about exactly that - control.
>
> **COMMENT:** This is a rational argument for how access to birth control is a win-win for young women, their partners, and the taxpaying public who might otherwise foot the welfare bill.

The choice to annotate the sentence as being COMMENTED_UPON or not depends on the question: how strong or obvious is the inference? The answer is ambiguous by nature and seems to partly depend on the annotator, given the number of total annotated sentences ranging from 123 to 212 (indicating that some annotators are more likely to annotate inference relations than others). Partly, however, it depends on the specific instance, indicated by the fact that all annotators had annotated multiple relations between sentences and comments that none of the others did.

The sentences that were not annotated at all (by none of the annotators and for none of the comments) typically included (personal) anecdotes or other background information to support or introduce the main arguments in the article. For example, the following four subsequent sentences introduce and illustrate the statements about the freeing powers of single-sex education that follow:

> **ARTICLE:** Years ago, during a classroom visit, I observed a small group of black and Latino high school boys sitting at their desks looking into handheld mirrors. They were tasked with answering the question, "What do you see?" One boy said, "I see an ugly face." Another said, "I see a big nose."

A major advantage of asking multiple annotators is that we can use different thresholds for selecting data. If we want to create a high quality set of clearly related sentences and comments, we can use only those sentences annotated by all. As suggested in Section 3, we can also select all sentences annotated by one or more person to aim for high recall. Nevertheless, this will not guarantee that no sentences are missed. Our results show that each additional annotator led to more candidate sentences, indicating that five annotators may be too few and new sentences would be added by a sixth annotator. If we want to find out how many relevant micro-proposition we miss, we can address this through a study where we apply the last two subtasks on complete texts and verify how many (dis)agreement pairs are missed in subtask 1.

5 Conclusion

We described a new task for argument mining based on our perspective framework and provided the results of a pilot annotation experiment aimed at identifying the sentences of an editorial article that are COMMENTED_UPON in a comment. Although a functional classification of statements was not part of our original goal, looking at argumentative texts from an interactive point of view did prove to shed new light on this more traditional argument mining task. Statements that are repeated, rephrased, attacked or supported by other debate participants seem to be the ones that are (at least perceived as) the main arguments of the text, especially when commented upon by multiple users. In contrast, statements that are not commented upon are likely to provide background information to support or introduce these arguments. We argued that annotator disagreement is not so much undesirable as it is insightful in tasks like this and reported on the distribution of the annotations. In our case, annotator disagreement appeared to be an indicator for the amount of inference that is needed to understand the relation between the sentence and the comment.

In the future, we plan to further experiment with the other two defined subtasks using a combination of expert annotation, semi-automatic approaches (textual similarity and entailment, generation of propositional relations) and crowdsourcing. Furthermore, we will include comment–comment relations (where one comment is a response to another) next to article–comment relations. The annotations and code for the experiment described in this paper are publicly available. [3]

References

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the 8th International Conference on Language Re-*

[3] github.com/ChantalvanSon/UnsharedTask-ArgumentMining-2016

sources and Evaluation (LREC 2012), pages 818–822, Istanbul, Turkey.

L. Aroyo and C. Welty. 2014. The three sides of CrowdTruth. *Human Computation*, 1:31–34.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 333–338, Istanbul, Turkey, May.

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.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, California, June. Association for Computational Linguistics.

Chantal van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo, and Piek Vossen. 2016. GRaSP: A multilayered annotation scheme for perspectives. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 812–817, Istanbul, Turkey.

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (ACL 2014)*, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Jie Yin, Paul Thomas, Nalin Narang, and Cecile Paris. 2012. Unifying local and global agreement and disagreement classification in online debates. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 61–69. Association for Computational Linguistics.