# A Shared Task on Argumentation Mining in Newspaper Editorials

**Johannes Kiesel** **Khalid Al-Khatib** **Matthias Hagen** **Benno Stein**
Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

## Abstract

This paper proposes a shared task for the identification of the argumentative structure in newspaper editorials. By the term "argumentative structure" we refer to the sequence of argumentative units in the text along with the relations between them. The main contribution is a large-scale dataset with more than 200 annotated editorials, which shall help argumentation mining researchers to evaluate and compare their systems in a standardized manner. The paper details how we model and manually identify argumentative structures in order to build this evaluation resource. Altogether, we consider the proposed task as a constructive step towards improving writing assistance systems and debating technologies.

## 1 Introduction

Even though argumentation theories have been studied extensively in many areas (e.g., philosophy), using these theories for mining real world text is a relatively new direction of research. Recently, argumentation mining has attracted many Natural Language Processing (NLP) researchers with papers published in major conferences and even a specialized workshop series.

Argumentation mining typically refers to the tasks of automatically identifying the argumentative units of a given text and the relations between them (i.e., support or attack). The automatic analysis of this discourse structure has several applications, such as supporting writing skills or assisting information-seeking users in constructing a solid personal standpoint on controversial topics.

To further foster the young field of argumentation mining, we propose a respective shared task to evaluate the current state of the art and compare to newly emerging ideas. According to the standard understanding of argumentation mining, we propose two focused sub-tasks: (1) unit identification and (2) relation extraction. The shared task will allow researchers to evaluate their systems in an open but standardized competition, which will help to push forward argumentation mining research.

For the corpus of the shared task, we are currently annotating a collection of newspaper editorials. These articles convey the opinions of their authors towards specific topics and try to persuade the readers of theses opinions. In order to do so, the authors support their opinions by reasons, which leads to an argumentative discourse. We plan to annotate at least 200 editorials from three different online newspapers paragraph-wise. Participants can use about two thirds of this corpus for training while the remainder will be used for evaluation.

The paper is structured as follows. Section 2 describes the argumentative structure that participants have to extract from the corpus that is described in detail in Section 3. Section 4 proposes the general scheme of the two sub-tasks while the task submission is outlined in Section 5. Conclusions are given in Section 6.

## 2 Argumentation Model

As the basis for the shared task, we employ a dialectical model of argumentation focusing on the conflict of opinions inspired by the definitions found in current research (Apothéloz et al., 1993; Bayer, 1999; Freeman, 2011; Stab and Gurevych, 2014), and especially that of Peldszus and Stede (2013).

The argumentation model consists of two elements: explicit argumentative units and implicit argumentative relations. Argumentative units are (explicitly written) text segments while argumentative relations correspond to inter-unit relationships (i.e., support or attack) that the reader implicitly establishes while comprehending the text. As a side remark, note that factual correctness is not modeled and also not part of our proposed shared task.

Although the argumentation model is primarily focused on dialectical discourse, it is also applicable to monologues in which the author switches roles. For instance authors of editorials often mention possible objections of others which they then attack when they switch back to their original role.

In applying the rather generic dialectical model, our proposed shared task is open for extensions/subtasks in many directions that are currently investigated in argumentation mining.

### 2.1 Detailed Model Description

An (argumentative) unit in our model is a consecutive segment of text that contains a formulation of at least one complete proposition which is written by the author to discuss, directly or indirectly, one of the main topics of the article.[1] Each proposition consists of an entity and a predicate that is assigned to this entity. For example, the unit “Alice is nasty” contains a predication of “nasty” to the entity “Alice,” but note that this also includes implicit entities or predicates like for instance in “He too.”[2]

An (argumentative) relation in our model is a directed link from one *base* unit to the *target* unit it *supports* or *attacks* most directly. In support relations, the base argues that the target is valid, relevant, or important. In this case, the base is also often referred to as premise or reason and the target as claim, conclusion or proposal (Mochales and Moens, 2011; Stab and Gurevych, 2014). In attack relations, the base argues that the target is invalid, irrelevant, or unimportant. Our proposed model only considers the most direct link (if any) for each base.

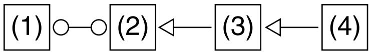

**Figure 1.** Structure of the example “Even though Eve says that [Alice is nasty]$^{(1)}$, I think [she is nice]$^{(2)}$. [She helps me a lot]$^{(3)}$. [She even taught me how to cook]$^{(4)}$!” Units are depicted as squares, supports by arrows, and an opposition by lines with a circle at each end.

For example, in Figure 1, even though Unit 4 also supports Unit 2, it is only linked to Unit 3 as it supports Unit 3 more directly.

The relations for a given text form one or more trees in our model; with the major claims of the text as their root nodes. Discussions in which a unit directly or indirectly supports or attacks itself can hence not be modeled.

Authors sometimes state the same proposition twice (*restatement*), or the directly contrary proposition if they take their opponents role (*opposition*). We take that into account by modeling special bidirectional support (for restatement) and attack (for opposition) relations. Note that in the case of restatements and oppositions, the tree structure of the text is no longer unambiguous. For example, in an equivalent structure to Figure 1, Unit 3 would attack Unit 1 instead of supporting Unit 2. In the proposed shared task, participants will have to identify restatements as supports and oppositions as attacks respectively, but all equivalent structures are scored as being correct (cf. Section 4 for further details).

### 2.2 Differences to Other Models

Our proposed model for the shared task does not explicitly categorize argumentative units into premises and claims since such a distinction gets problematic when claims are premises for further claims. Stab and Gurevych (2014) try to handle this problem by introducing so-called major claims—which are supported by claims. However, for longer reasoning chains, in which even these major claims would support further claims, this approach fails. In our model, premises and claims are defined relative to each other by support relations: every base is a premise to the target claim it supports. In this way, we can adequately represent reasoning chains of any length.

Although more fine-grained labels, such as different types of attack and support, will be annotated in our corpus, we will drop this distinction for the

[1] For editorials, the title often introduces (one of) its main topics.

[2] Questions can also be formulations of propositions if the answer is already suggested (i.e., rhetorical questions).

task in order to reduce its complexity as well as have a more straightforward evaluation (cf. Section 4 for more details). In comparison to the model of Peldszus and Stede (2013), our model employed in the shared task will subsume the types “basic argument” and “linked support” under support and “rebutting” and “undercutting” under attack.

Unlike Peldszus and Stede (2013), we do not directly distinguish between units from proponent or opponent views since this distinction is difficult for discussions evolving around several topics. In our model, such a distinction is present on a local level: when one unit attacks another.

## 3 Corpus

In order to acquire high quality opinionated articles, we only consider editorials from newspaper portals which include a separate section for opinion articles and have a high international reputation. For our corpus, we selected the portals of Al Jazeera, Fox News, and The Guardian. From their opinion section we crawled 1415, 1854, and 110 articles, respectively. From each crawl, we exclude particularly short or long articles and select the 75 articles with the most comments. We see the number of comments as an indicator of how controversial the discussed topic is and expect articles with many comments to contain more conflicting arguments.

After the editorials are selected, they are annotated based on our model (cf. Section 2). The annotation process is conducted with three workers from the online platform oDesk.[3] We first annotate ten articles in a pilot study and annotate the remaining ones (or even more) after we inspected the results. Annotation will be carried out in three steps: the identification of (1) the topics, (2) the argumentative units, and (3) the argumentative relations. After each step, the annotations of the workers are manually unified to create a single consistent annotation as foundation for the next step.

## 4 Task Description

The task of argumentative structure extraction can be divided into two steps: the identification of argumentative units and the identification of the relations between them. Accordingly, we propose two sub-tasks focusing on one of these steps each.

### 4.1 Argumentative Unit Classification

For each article in the corpus, the participants get a list of the main topics and a list of propositions that they have to classify as argumentative with respect to one of the given topics or not.

Since this is a binary classification task, standard accuracy is an appropriate measure for evaluation purposes. Let $P$ be the set of propositions in the corpus, $\mathrm{c}_S(p)$ be the system’s predicted class (argumentative or not) for proposition $p$, and $\mathrm{c}_G(p)$ be the gold standard class of the proposition $p$. Moreover, let $\mathrm{C}_G(p, c)$ be 1 if $\mathrm{c}_G(p) = c$ and 0 otherwise. The unit classification accuracy of a system $S$ then is:

$$\mathrm{accuracy}_{\mathrm{C}_G}(S) = \frac{\sum_{p \in P} \mathrm{C}_G(p, \mathrm{c}_S(p))}{|P|},$$

where $|P|$ is the number of propositions.

The participants’ results will be compared to several baselines, one natural one being the random guessing of a class based on the class distribution in the training set.

### 4.2 Argumentative Relation Identification

For each paragraph in the corpus, the participants get the text and the argumentative units as input and have to produce the support and attack relations between the units in the paragraph. The relations extracted from each paragraph have to form one or more trees with the units as nodes. Furthermore, relations always have to be directed towards the root of the tree. If several structures are possible for one paragraph due to restatements and oppositions (cf. Section 2), any of them will get a perfect score. If a restatement (or opposition) occurs in the test corpus, systems are expected to produce a support (or attack) relation between the units in any direction.

This sub-task uses unit-wise accuracy as evaluation measure. Our proposed model states that each base can have only one target (cf. Section 2). The same restriction applies to the systems of the participants. This allows us to define unit-wise accuracy as follows. Let $U$ be the set of units in the corpus and let $\mathrm{r}_S(u)$ be the relation with unit $u$ as a base in the system output or a special no-relation-symbol $\perp$

[3] https://www.odesk.com/

if no such relation exists. Furthermore, let $\mathrm{R}_G(u, r)$ be 1 if $r$ is a correct relation with base $u$ with regard to the gold standard and 0 otherwise. The relation identification accuracy of a system $S$ then is:

$$\text{accuracy}_{\mathrm{R}_G}(S) = \frac{\sum_{u \in U} \mathrm{R}_G(u, \mathrm{r}_S(u))}{|U|},$$

where $|U|$ is the number of units in the corpus. A relation $r$ with base $u$ is correct if the same or an equivalent relation with regard to polarity (support/attack) and target unit exists in the gold standard. Here, equivalence takes into account restatements and oppositions (cf. Section 2). Moreover, if $r$ is $\perp$, then $\mathrm{R}_G(u, r)$ is 1 if and only if there is also no relation with $u$ as a base in the gold standard.

Similar to the first sub-task, we plan to compare the results of the participants to simple baseline approaches. One such baseline is random guessing according to the distributions in the training set. Another approach is a classifier which uses only one feature (e.g., the output of a textual entailment software package).

## 5 Submission

For the participants' submissions, we want to employ recent advances in reproducible computer science and ask the participants to submit their software instead of submitting their results on the test dataset. In detail, the participants will setup their systems with a unified interface on a remote virtual machine. This machine is then used to evaluate the systems, which makes the experiments directly reproducible in the future.

System submissions are currently becoming increasingly popular in shared tasks. For example, the CoNLL 2015 shared task on shallow discourse parsing[4] applies this technology. We plan to use the same system as the CoNLL task, TIRA (Gollub et al., 2012),[5] which is already successfully applied in the PAN workshops on plagiarism detection.[6]

## 6 Conclusions

We propose a shared task for mining the argumentative structure in newspaper editorials. This includes modeling the argumentative discourse, creating an annotated corpus, and proposing two sub-tasks for automatic argumentation mining. The sub-tasks are the identification of argumentative units in a text and the identification of relations between the units. We propose appropriate evaluation measures, and suggest to use a new submission approach to increase the reproducibility of the participants' systems.

We believe that it is of great importance for the further development of argumentation mining to establish a shared task in which different systems are evaluated against each other in a standardized and objective manner. Any comments and requests from the research community can still be included in the final task design.

[4] http://www.cs.brandeis.edu/~clp/conll15st/

[5] http://www.tira.io/

[6] http://www.pan.webis.de

## References

Denis Apothéloz, Pierre-Yves Brandt, and Gustavo Quiroz. 1993. The Function of Negation in Argumentation. *Journal of Pragmatics*, 19(1):23–38.

Klaus Bayer. 1999. *Argument und Argumentation: logische Grundlagen der Argumentationsanalyse*, volume 1 of *Studienbücher zur Linguistik*. Westdeutscher Verlag.

J. B. Freeman. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer.

Tim Gollub, Benno Stein, and Steven Burrows. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation Mining. *Artificial Intelligence and Law*, 19(1):1–22.

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.

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, August.