Exploring Crowdsourceable Annotation Protocol for Argumentation Schemes

Naoya Inoue[1,2,a)] Paul Reisert[2,b)] Kentaro Inui[1,2,c)]

Abstract: Identifying the type of arguments (i.e. *argumentation scheme* or *AS*) made in argumentative texts is an important core technology for detecting implicit premises. As of yet, however, there is no reliable, large-scale corpus annotated with ASs. This paper explores the potential of crowdsourcing as a plausible way for creating such an annotated corpus of ASs. To make the annotation task easier, we carefully simplify its annotation protocol. This allows us to outsource the annotation task to non-expert workers, while maintaining the annotation target to everyday, frequently observed arguments. This paper shows how to design such a crowdsourcing task, and our detailed analysis of the results demonstrates that the annotation task is reasonably achieved by the workers. We plan to publicly release an annotated corpus for research purposes.

1. Introduction

Everyday arguments are usually *enthymemes* [1], or *incomplete* in the sense that all premises to draw a conclusion are not presented [35]. Consider the following argument:

(1) *Japan should invest more in space exploration. Space technologies such as satellites are expected to be advanced.*

Analyzing the argument in detail, the logic can be summarized as follows:

- If the action "*more investment in space exploration*" is brought about, the consequence "*the advancement of space technologies*" will occur.
- The consequence is desirable.
- Therefore, the action should be brought about.

In Example (1), however, premises such as (i) the causality between the action and its consequence (i.e. *investment in space exploration promotes space technologies*), and (ii) the writer's value judgment towards the consequence (i.e. *advancement of space technology is desirable*), are implicit. Detecting such implicit premises is a long-standing goal in artificial intelligence. Applications such as automated essay scoring and critical thinking training in educational context would be benefited from this technology [29], [34].

A study on enthymemes has been received much attention [2], [6], [11], [24] in the field of argument mining [15]. In the literature, the implicit premise identification task has typically been formalized as an argumentation pattern identification task. Feng et al. [5] develop a machine learning classifier to recognize the type of an argument made in an argumentative text within five types of Walton [35]'s Argumentation Schemes (ASs). An AS is essentially a typical reasoning form consisting of a set of premises with slots (e.g. "If *A* is brought about, good consequence *C* will occur.") and a conclusion (e.g. "*A* should be brought about") (see Sec. 3.1 for further details). Identifying an AS and filling the slots explicates the important premises (explicitly or implicitly) assumed in an argument, which would help a machine detect what premises are potentially left implicit.

Razuvayevskaya et al. [25] conduct a feasibility study on designing the task of implicit premise identification for *a fortiori* arguments [28]. They focus on arguments represented by the linguistic pattern "*X, let alone Y*", which always presupposes an implicit premise that *X* has some relative relation with *Y* (e.g. "*I've never been to Germany, let alone Europe*" presupposes that *Germany* is a part of *Europe*). They manually define 11 semantic relations as an inventory of argumentation patterns and aim at identifying implicit premises via the relation classification task.

As described in Sec. 2, another research direction includes annotating with implicit premises in a natural language form [3], [8]. However, those predefined pattern-based formalizations have an advantage in its machine-friendly, structured output like template-based information extraction, which would be more usable for downstream applications.

Nevertheless, there is no reliable, large-scale corpus annotated with predefined schemes on a wide range of arguments. The AraucariaDB corpus [26] used by Feng et al. [5] contains 664 argumentative texts extracted from news articles etc. which are manually annotated with ASs (from approximately 30 ASs). However, the inter-annotator agreement (IAA) is not reported. In addition, an IAA of AS annotation is proven to be low in follow

[1] Tohoku University, 6-6-05 Aramaki Aza Aoba, Aobaku, Sendai, Miyagi 980-8579, Japan
[2] RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
[a)] naoya-i@ecei.tohoku.ac.jp
[b)] paul.reisert@riken.jp
[c)] inui@ecei.tohoku.ac.jp

Table 1 Summary of corpora annotated with implicit premises.

Study	Target Arg.	Representation	Data size	IAA
AraucariaDB [26]	Claim & Premise	30 Walton's Arg. Schemes [35][*1]	660	N/A or Low [14]
Razuvayevskaya et al. [25]	X let alone Y	11 semantic relations	100	High
Harbenal et al. [8]	Claim & Premise	Natural Language	2,000	N/A
Boltuzic et al. [3]	Claim & Premise	Natural Language	625	N/A

up studies [14], [18]. Razuvayevskaya et al. [25] report a high IAA, but they consider only a fortiori arguments. The key lessons from the previous work is that the more argumentation schemes we have, the harder the annotation task is. We believe that the challenge would be how to make a trade off between a variety of annotated argumentation patterns, the coverage of arguments and annotation reliability.

This paper explores the potential of crowdsourcing as a plausible way for creating a reliable, large-scale corpus annotated with predefined ASs. Given the complexity of the annotation task, one might think that the annotation task cannot be reasonably achieved by non-expert workers. We hypothesize, however, that with a proper simplification and instruction of the annotation task, we are able to obtain reasonable annotation results from qualified, non-expert workers.

Specifically, following [25], we focus on specific argument types—*policy arguments* where the main claim insists that some action should or should not be brought about. We then annotate these arguments with three types of ASs, namely a variant of Argument from Consequences (AfC) [35], which makes an argument that one should do some action based on the desirability of action's consequence (e.g. Example (1)). While focusing on arguments frequently observed in everyday arguments [21][*2], these restrictions enable the annotation task to be much simpler, thereby making the task outsourceable to crowdsourcing. The contribution of this paper is (i) to show how to design the crowdsourcing task, and (ii) to demonstrate that the annotation task is reasonably achieved by the crowdsource workers through a detailed analysis of the crowdsourcing results. We also plan to publicly release an annotated corpus for research purposes.

2. Related work

Recognizing argumentative structures in unstructured texts is an important task for many NLP applications. Argument mining is an emerging, leading field of argumentative structure identification in the NLP community [16]. It involves a wide variety of subtasks for argumentative structure identification such as argument component identification/classification [27], [30], stance classification [10], [22] and argumentative relation detection [4], [19], [20]. Previous studies on discourse analysis also explored discourse theories to structuralize a text such as Rhetorical Structure Theory [17] and Segmented Discourse Representation Theory [13]. These tasks has been useful for applications such as essay scoring, document summarization, etc. [31].

On the other hand, less attention has been paid to the task of identifying implicit premises. The research community has not reached a consensus on the task design of implicit premise identification yet and several task designs have been proposed. Table 1 summarizes such previous efforts and the linguistic resources. In the rest of this section, we briefly review each previous work and then give a discussion.

2.1 Predefined pattern-based approaches

The AraucariaDB corpus [26] is a pioneering work on implicit premise identification and partially employed by Feng et al. [5]. This corpus has 660 arguments from a wide variety of sources (news articles etc.), which are manually annotated either with Walton's AS [35], Katzv's AS [12] or Pollock's AS [23]. On top of ASs, the implicitness of each premise is also annotated. Whereas the AraucariaDB corpus [26] covers a wide range of arguments and ASs, the corpus is not reliable because the inter-annotator agreement (IAA) is not reported. In addition, an IAA of Walton's AS annotation is proven to be low in follow up studies [14], [18].

Razuvayevskaya et al. [25] conduct a feasibility study on the task of implicit premise identification for *a fortiori* arguments [28]. Specifically, given two propositions P_1, P_2, an a fortiori argument has two premises: (i) P_1 semantically subsumes P_2 and (ii) P_1 is true. It then concludes that P_2 is true as well. For example, given that *John likes mathematics*, with an a fortiori argument, one would conclude that *John passes a mathematics exam*, assuming that *to like something X* entails *to pass an exam on X*. In real-life arguments, premise (i) is often left implicit. Razuvayevskaya et al. [25] focus on *a fortiori* arguments represented by the linguistic pattern "*X, let alone Y*". This pattern always presupposes an implicit premise that *X* has some semantic relation with *Y*. For example, "*I've never been to Germany, let alone Europe*" presupposes that *Germany* is a part of *Europe*. They cast an implicit premise identification task as identifying how P_1 is semantically subsumed by P_2. They manually define 11 semantic relations (e.g. is-a-part-of, entails). To examine the feasibility of the annotation task, they randomly extracted 100 sentences from the BNC corpus and obtained a Cohen's Kappa of 0.75 ("good agreement"). This task design demonstrates a good IAA; however, their annotation scheme targets only *let alone* constructions.

This study falls into the predefined pattern-based approach. As discussed in Sec. 1, this approach has an advantage in its machine-friendly, structured output like template-based information extraction, which would be more usable for downstream applications.

2.2 Natural language-based approaches

Habernal et al. [8] create a binary classification task for ar-

[*1] The arguments are also annotated with 25 Katzv's Arg. Schemes [12], or 7 Pollock's Arg. Schemes[23].
[*2] http://idebate.org/

gument comprehension.[*3] Given a claim C (e.g. *scholarships would give women a chance to study*), reason R (e.g. *scholarships would take women from the home*), and two warrants W_1 (e.g. *Miss America gives honors and education scholarships.*), W_2 (e.g. *Miss America is good for women*), the task is to choose a correct warrant which explains why R makes sense as a reason for C (i.e. W_1 in this case). Using crowdsourcing, they create approximately 2,000 (claim, reason, warrant, alternative warrant) triples for 188 topics.

Boltuzic et al. [3] annotate 625 claim-support pairs provided by [10] for 5 topics with implicit premises in a natural language form.[*4] To annotate implicit premises, they provided the following instruction "we ask the annotators to provide the premises that bridge the gap between the two claims" to the annotators.

Although these two studies established a scalable way for annotating implicit premises, the representation of an implicit premise is in a natural language form. The disadvantage of a natural language form is that the quality assurance is difficult, because the IAA cannot be calculated. Additionally, the output would be less usable for downstream applications.

3. Annotating policy arguments with Argument from Consequences

3.1 Annotation principle

In a text, there can be multiple, potentially nested, ASs [35], and identifying all of them would be a difficult task. Indeed, Musi et al. [18] report that the inter-annotator agreement of annotating Walton [35]'s ASs results in Fleiss's $\kappa = 0.31$ ("fair agreement") even if the annotators are trained and only a subset (8 types) of schemes are annotated. Therefore, this work has three restrictions. First, we focus only on policy arguments. Second, we locally annotate a *pair* of argumentative components (i.e. a claim and a premise) with one scheme.

Third, we restrict an inventory of ASs to only three schemes related to AfC, which makes an argument that one should do some action based on the desirability of action's consequence. Walton [35]'s Argumentation Scheme theory introduces more than 50 ASs, which describe common reasoning patterns (i.e. ASs) in everyday arguments. Each AS consists of (i) a set of premises, (ii) a conclusion[*5] and (iii) critical questions that assess an argument's acceptability. The original AS theory describes more than 50 ASs, but policy arguments are commonly done by a variant of AfC [5], [35]. In addition, the AfC is expected to be more comprehensible to non-expert crowdworkers than other ASs because of its usage in daily conversations.

3.2 Annotation protocol

Given an argument (i.e. a claim and a premise), our goal is to annotate the type of AS using the list shown in Table 2. For simplicity, we assume that a premise always corresponds to C. For example, we would annotate Example (1) with Argument from Positive Consequence (AfPC), where A = "*Japan invest more in space exploration*" and C = "*Space technologies such as satellites are expected to be advanced.*". On the other hand, we annotate the following example with Other, a special scheme indicating that none of the ASs in Table 2 are applicable:

Table 2 Argumentation Schemes used in this study.

Argument from Positive (Negative) Consequence (AfPC (AfNC))	
Premise 1:	If action A is (not) brought about, consequence C will occur.
Premise 2:	C is desirable (undesirable).
Conclusion:	A should (not) be brought about.
Prudential Argument from Negative Consequence (PruAfNC)	
Premise 1:	If action A is not brought about, consequence C will occur.
Premise 2:	C is undesirable.
Conclusion:	A should be brought about.

(2) *Japan should invest more in space exploration. Space exploration is ambitious.*

The premise does *not* mention a consequence of an action *but* the attribute of *space exploration*.

The perfect annotation scheme would be to annotate what type of premise is presented in an argument, in addition to the type of AS. However, we leave it to future work, as this is the first attempt on using crowdsourcing for annotating ASs. The perfect annotation would tell us, for instance, that Example (1) implicitly assumes (i) the causality between *investment in space exploration* and *advancement of space technologies* (by **Premise 1**), and (ii) the desirability of the consequence *advancement of space technologies* (by **Premise 2**). A downstream application can use this kind of information for predicting its output. For example, an automated essay scorer (AES) might be able to learn that the implicitness of desirability of *advancement of space technologies* results in a low score, but the implicitness of desirability of *increase of brutal crimes* does not (because of its obviousness).

Furthermore, an AES could advise the user to critically think whether the implicit premises are really true and rethink the logic of arguments. Interestingly, it is controversial whether the advancement of space technologies is good or not, as it would promote *a diversion to military use*. Furthermore, *satellites* are not examples of *space exploration* (i.e. those of terrestrial technology).

3.3 Dataset construction via crowdsourcing

The main purpose of this work is to create a scalable way for constructing an reliable, annotated corpus of ASs. Inspired by Habernal et al. [8], who create a benchmark dataset for discriminating a correct implicit premise from wrong one via crowdsourcing, we designed a crowdsourcing task for the annotation scheme designed in Sec. 3.1 and 3.2.

To make the annotation task easier, we formulated the task as a *sentence verification task*. We first show workers an argumentative text with its main claim (as the title) and specify a premise to annotate. We then show sentences to be verified, which reveal an AS underlying the main claim and the premise.

To automatically generate sentences, we use the following linguistic templates. Let C be an affirmative form of main claim (e.g. *Germany were to introduce the death penalty*), NC be a negative form of main claim (e.g. *Germany does not introduce the death*

[*3] https://competitions.codalab.org/competitions/17327
[*4] http://takelab.fer.hr/data/argpremises/
[*5] The terms "conclusion" and 'claim' are used interchangeably.

Fig. 1 Crowdsourcing annotation interface.

penalty), and P be a premise (e.g. *Death penalty convicts innocent people*). The templates are the following:

- AfPC: "*If C, a good consequence will occur. The good consequence is that P.*"
- PruAfNC: "*If NC, a bad consequence will occur. The bad consequence is that P.*"
- AfNC: "*If C, a bad consequence will occur. The bad consequence is that P.*"

Fig.1 illustrates an actual screen shown to the workers, where the writer's argument is: "if '*the fine is not increased*', '*one can hardly...*', a bad consequence, will occur. Therefore, the *fine should be increased*." The worker needs to select "No" to the first question, whereas the worker needs to select "Yes" to the second question.

In some cases, since a premise is automatically extracted from a text (see Sec. 4.1), generated sentences are not always grammatically correct but still capture the writer's idea (e.g. *If the fine is not increased, a bad consequence will occur. The bad consequence is that and you step into one of the many 'land mines'.*) We thus instructed workers to select "Yes" when the sentence without adjunct phrases "For one thing," etc. makes sense.

4. Experiments

4.1 Setting

Dataset. We use the argumentative micro text corpus [21][*6], a small, but reliable argumentative corpus. The corpus contains 89 policy arguments (out of 112), each consisting of roughly five segments composed of a topic question (23 types; e.g. *Should video games be made olympic?*), a main claim, and premises (see Fig. 1 for an example). We extracted 164 premises that directly support a main claim based on the annotated support relations (e.g. (1) and (2) in Fig. 1). For assisting the workers with understanding the stance of the text, we converted each topic question to a title statement consistent with the main claim (e.g. see "Title" in Fig. 1).

[*6] https://github.com/peldszus/arg-microtexts

Table 3 Distribution of inter-annotator agreements (should arguments).

Split	Partial	Perfect
0.012 (1/86)	0.430 (37/86)	0.558 (48/86)

Table 4 Distribution of inter-annotator agreements (should-not arguments).

Split	Partial	Perfect
0.000 (0/78)	0.385 (30/78)	0.615 (48/78)

Table 5 Distribution of argumentation schemes.

Should Args.		**Should Not Args.**	
Scheme	Portion	Scheme	Portion
AfPC	.314 (27/86)	AfNC	.333 (26/78)
PruAfNC	.140 (12/86)	Other	.667 (52/78)
Other	.547 (47/86)		

CrowdSourcing. We use Figure Eight[*7] as a crowdsourcing platform. Per one premise, we hired three workers and paid five cents as a reward. To assure the quality of annotation, we prepared gold-standard answers for 15 premises (henceforth, *test question*). The test questions are then randomly placed into the crowdsourcing task, where the workers do not know whether the instance that they are working on is a test question or not. The workers are required to maintain an accuracy of 70% or more on these test questions (henceforth, *qualified workers*); otherwise, the annotation results are discarded and they are not paid.[*8]

4.2 Results

Table 3 and 4 show the agreement between workers for should arguments (i.e. policy arguments with "should do" claim) and should-not arguments (i.e. policy arguments with "should not do" claim), where Split, Partial, and Exact indicates the number of instances where all three workers had different labels, two workers had the same labels, and all three workers had the same label, respectively. The results indicate that the annotation task is reasonably done by non-expert, qualified workers for both types of policy arguments.

Table 5 shows the distribution of ASs, where the final labels are obtained by majority vote. The results indicate that a roughly half of the should and should not arguments (45.4% and 33.3%, respectively) are captured by AfC. For example, three workers labeled the following instance as AfNC:

(3) ***Claim***: *Germany should not introduce the death penalty.*
Premise: *Capital punishment will not deter anyone else from an atrocity. (*`micro_b031`[*9]*)*

where the writer cites the undesirable consequence of *capital punishment* (i.e. the proposed action in debate), namely it will not work as a deterrent against brutal crimes.

Table 6 Example Other instances.

Premise Type	Portion	Claim	Premise
RightOrObligation	25.0%	*Germany should not introduce the death penalty*	*A door must remain open for making amends.*
AdditionalInfo	22.1%	*Germany should not introduce the death penalty*	*Courts are also subject to human error.*
ProblemToBeSolved	14.7%	*The statutory retirement age should not remain at 63 years in the future*	*Thus social security and pension costs are increasing.*
Direct Rebuttal	7.4%	*Germany should not introduce the death penalty*	*Handing out capital punishment is unethical.*
Satisfactory	4.4%	*There should not be a cap on rent increases for a change of tenant*	*Rent prices are already regulated in favour of tenants due to existing laws and the rent index.*
Alternative	4.4%	*All universities in Germany should not charge tuition fees*	*The anticipated objectives of tuition fees can be achieved by other means.*
Expert	2.9%	*Public health insurance should cover treatments in complementary and alternative medicine*	*Besides many general practitioners offer such counseling and treatments in parallel anyway*
AFortiori	2.9%	*The fine for leaving dog excrements on sideways should be increased*	*Besides you're not allowed to leave other rubbish without punishment.*
Misc.	16.2%	-	-

4.3 Analysis

4.3.1 Other instances

To observe what types of arguments are categorized as Other, we manually analyzed such instances. We randomly selected 10 topic questions and analyzed 78 Other instances under these topics. We found that 11.7% of them were misclassified AfC instances. We manually classified the remaining 88.3% of Other instances according to the type of premise, which is shown in Table 6. To summarize, most of Other instances are still AfC, but the premise did not explicitly say an action's consequence. We elaborate on the most frequent three premise types below.

RightOrObligation. The most frequent patterns make an argument appealing to public rights or obligations (or needs). For example, for the RightOrObligation instance in Table 6, the argument claims a human right that *a door must remain open for making amends*, even though people commit a crime. The argument then implicitly states that the death penalty violates this right and therefore must not be introduced. Note that, in the original argument, this logic is left implicit and only the right is explicated. However, the argument can be still considered AfC, because it is on the grounds that the violation of such a right is an undesirable consequence. Another example includes the claim "*All universities in Germany should not charge tuition fees*" and the premise "*Every German citizen has a right to education.*"

Another example is that the argument states a need, obligation or ideal situation and then implicitly claims that the proposed action satisfies it. Consider the following example:

(4) ***Claim***: *Shopping malls should generally be allowed to open on holidays and Sundays*
Premise: *Plus, the state wants me to spend my money, (micro_b015)*

where the need is stated in **Premise**. The argument implicitly states that opening shopping malls on holidays and Sundays would satisfy this need, namely promote people to spend their money, which is a desirable consequence. Another example includes the claim "*The Berlin Tegel airport should not remain operational after the opening of the Berlin Brandenburg airport*" and the premise "*Also, Berlin urgently needs inner-city areas for business and industry, and for housing.*".

AdditionalInfo. Premises in this pattern simply add information such that the likelihood of desirable (or undesirable) consequences occurring gets higher. For the AdditionalInfo instance in Table 6, the belief that *courts are subject to human error* makes the likelihood of undesirable consequence much higher. These arguments are still AfC, because they appeal to an action's consequences. Another example includes the claim "*Public health insurance should cover treatments in complementary and alternative medicine*" and the premise "*Often natural treatments do not put as much of a strain on the patient.*"

ProblemToBeSolved. An argument mentions a current undesirable situation, and then implicitly states that the proposed action makes the situation better (or even worse). For the ProblemToBeSolved instance in Table 6, the premise mentions an undesirable current situation, namely *increase of social security and pension costs*. The argument implicitly states that this undesirable situation should be improved by a proposed action, namely *the change of retirement age*. Similarly to the above patterns, this type of argument is still AfC, but does not explicate its consequence and value judgement. Another example includes the claim "*All universities in Germany should not charge tuition fees*" and the premise "*Even without tuition fees half of the student population has to work while studying and hence has less time for studying or recreation.*", where the argument is that the current undesirable situation will be made even worse by the action.

In future work, we will extend our annotation scheme to capture these majority patterns.

4.3.2 Towards automation and large-scale annotation

Towards automating AS identification, we manually analyzed the annotation results. Some instances have cue phrases evoking AfC, such as "*result in*" and "*would*", but most of the other instances do not. In future work, we will leverage an external knowledge base to recognize the causality between a main claim and a premise. For example, consider the following example labeled as AfNC:

(5) ***Claim***: *Germany should not introduce the death penalty.*
Premise: *Moreover it turns out time and again that innocent people are also convicted and executed. (micro_b027)*

The world knowledge that "*introduction of the death penalty*" could cause "*conviction of innocent people*" would help a ma-

[*7] https://www.figure-eight.com/
[*8] It took 12 hours to complete the crowdsourcing task.
[*9] This id refers to the argument in the argumentative micro text corpus.

chine recognize this as the AfNC.

We also observed that given a topic question, diversity of world knowledge and patterns of argument are almost closed. For example, in the topic question "*All universities in Germany should not charge tuition fees*", most of the argument appeal to the violation of educational rights (i.e. RightOrObligation in Table 6) even though the linguistic expressions in their premises are different (e.g. *Every German citizen has a right to education*, *Studying and taking higher degrees must remain a basic right for everyone*). In future work, we plan to design a micro-domain argumentation analysis task under specific, two or three debate topics. We will largely expand our annotation under several topics and then formulate the task of argumentation analysis as the task of grounding arguments on a set of short passages encoding debate topic-specific world knowledge, similarly to the Entity Linking task.

5. Conclusion

We have explored the potential of crowdsourcing as a plausible way for creating a reliable, large-scale corpus annotated with ASs. To make the annotation task easier, we have carefully simplified its annotation protocol by limiting targeted arguments to policy arguments and an inventory of ASs to a variant of AfC. As a result, the annotation task has become outsourceable to non-expert crowdworkers, while focusing on a wide range of real-life arguments. Our detailed analysis of the crowdsourcing results has demonstrated that the annotation task is reasonably achieved by crowdworkers.

Based on the proposed crowdsourcing task, we plan to expand our annotation to other argumentative corpora [7], [9], [32], [33], etc. and publicly release the annotated corpus for research purposes. We also plan to create a computational model to assess the difficulty of this task. Future work includes designing a crowdsourcing task for annotating the implicitness of premises and developing a sophisticated, automated AS prediction model using an external knowledge base.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR1513, Japan.

References

[1] Aristotle and Roberts, W. R.: *Aristotle: Rhetoric*, Modern Library (1954).

[2] Black, E. and Hunter, A.: A relevance-theoretic framework for constructing and deconstructing enthymemes, *Journal of Logic and Computation*, Vol. 22, No. 1, pp. 55–78 (online), DOI: 10.1093/logcom/exp064 (2012).

[3] Boltuzic, F. and Šnajder, J.: Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates, *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pp. 124–133 (2016).

[4] Cocarascu, O. and Toni, F.: Identifying attack and support argumentative relations using deep learning, pp. 1385–1390 (online), available from ⟨http://www.aclweb.org/anthology/D17-1145⟩ (2017).

[5] Feng, V. W. and Hirst, G.: Classifying Arguments by Scheme, *Computational Linguistics*, pp. 987–996 (online), available from ⟨http://ftp.cs.toronto.edu/pub/gh/Feng+Hirst-2011.pdf⟩ (2011).

[6] Green, N. L.: Identifying Argumentation Schemes in Genetics Research Articles, *Proceedings of the 2nd Workshop on Argumentation Mining*, pp. 12–21 (online), available from ⟨http://aclweb.org/anthology/W/W15/W15-0502.pdf⟩ (2015).

[7] Habernal, I. and Gurevych, I.: What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation, pp. 1214–1223 (online), available from ⟨https://aclweb.org/anthology/D/D16/D16-1129.pdf⟩ (2016).

[8] Habernal, I., Wachsmuth, H., Gurevych, I. and Stein, B.: The Argument Reasoning Comprehension Task, (online), available from ⟨http://arxiv.org/abs/1708.01425⟩ (2017).

[9] Hasan, K. S. and Ng, V.: Stance Classification of Ideological Debates : Data , Models , Features , and Constraints, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, No. October, pp. 1348–1356 (online), available from ⟨http://www.aclweb.org/anthology/I13-1191⟩ (2013).

[10] Hasan, K. S. and Ng, V.: Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 751–762 (2014).

[11] Hosseini, S. A., Modgil, S. and Rodrigues, O.: Enthymeme Construction in Dialogues using Shared Knowledge, *Frontiers in Artificial Intelligence and Applications*, Vol. 266, pp. 325–332 (online), DOI: 10.3233/978-1-61499-436-7-325 (2014).

[12] Katzav, J. and Reed, C.: A classification system for arguments, Technical report, Department of Applied Computing, University of Dundee (2004).

[13] Lascarides, A. and Asher, N.: Segmented discourse representation theory: Dynamic semantics with discourse structure, *Computing meaning*, Vol. 3, p. 87124 (online), available from ⟨http://www.springerlink.com/index/m9856476613455n5.pdf⟩ (2007).

[14] Lawrence, J. and Reed, C.: Argument Mining using Argumentation Scheme Structures, *Proceedings of the 6th International Conference on Computational Models of Argument (COMMA 2016)*, Vol. 0, pp. 379 – 390 (online), DOI: 10.3233/978-1-61499-686-6-379 (2016).

[15] Lippi, M. and Torroni, P.: A Argumentation Mining: State of the Art and Emerging Trends, *ACM Transactions on Internet Technology*, Vol. V, pp. 1–25 (online), DOI: 10.1145/0000000.0000000 (2015).

[16] Lippi, M. and Torroni, P.: Argument mining: A machine learning perspective, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9524, pp. 163–176 (online), DOI: 10.1007/978-3-319-28460-6_10 (2015).

[17] Mann, W. C. and Thompson, S. A.: Rhetorical Structure Theory: A Theory of Text Organization, Technical report (1987).

[18] Musi, E., Ghosh, D. and Muresan, S.: Towards Feasible Guidelines for the Annotation of Argument Schemes, *Third Workshop on Argument Mining (ArgMining2016)*, Vol. 30, No. 3, pp. 82–93 (2016).

[19] Niculae, V., Park, J. and Cardie, C.: Argument Mining with Structured SVMs and RNNs, (online), available from ⟨http://arxiv.org/abs/1704.06869⟩ (2017).

[20] Peldszus, A. and Stede, M.: Joint prediction in MST-style discourse parsing for argumentation mining, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, No. September, pp. 938–948 (online), available from ⟨http://aclanthology.info/papers/joint-prediction-in-mst-style-discourse-parsing-for-argumentation-mining⟩ (2015).

[21] Peldszus, A. and Stede, M.: An Annotated Corpus of Argumentative Microtexts, *Studies in Logic and Argumentation*, (online), available from ⟨http://www.ling.uni-potsdam.de/ peldszus/eca2015-preprint.pdf⟩ (2016).

[22] Persing, I. and Ng, V.: Modeling Stance in Student Essays, pp. 2174–2184 (online), available from ⟨https://www.aclweb.org/anthology/P/P16/P16-1205.pdf⟩ (2016).

[23] Pollock, J. L.: *Cognitive carpentry: A blueprint for how to build a person*, Mit Press (1995).

[24] Rajendran, P. and Parsons, S.: Contextual stance classification of opinions : A step towards enthymeme reconstruction in online reviews, pp. 31–39 (online), available from ⟨http://www.aclweb.org/anthology/W16-2804⟩ (2016).

[25] Razuvayevskaya, O. and Teufel, S.: Finding enthymemes in real-world texts : A feasibility study, Vol. 8, pp. 113–129 (online), DOI: 10.3233/AAC-170020 (2017).

[26] Reed, C.: Araucaria: Software for argument analysis, diagramming and representation, *International Journal of AI Tools*, Vol. 13, No. 4, pp. 961–980 (online), DOI: 10.1142/S0218213004001922 (2004).

[27] Reed, C. and Palau, R. M.: Language resources for studying argument, *Proceedings of the 6th conference on language resources and evaluation*, pp. 91–100 (online), available from ⟨https://lirias.kuleuven.be/handle/123456789/197357⟩ (2008).

[28] Sidgwick, A.: The A Fortiori Argument, *Mind*, Vol. 25, No. 100, pp. 518–521 (online), available from ⟨http://www.jstor.org/stable/2248858⟩ (1916).

[29] Song, Y., Heilman, M., Beigman Klebanov, B. and Deane, P.: Ap-

plying Argumentation Schemes for Essay Scoring, *Proceedings of the First Workshop on Argumentation Mining*, No. 2011, pp. 69–78 (online), available from ⟨http://www.aclweb.org/anthology/W/W14/W14-2110⟩ (2014).

[30] Stab, C. and Gurevych, I.: Annotating Argument Components and Relations in Persuasive Essays, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1501–1510 (online), available from ⟨http://www.aclweb.org/anthology/C14-1142⟩ (2014).

[31] Stab, C. and Gurevych, I.: Parsing Argumentation Structures in Persuasive Essays, *arXiv preprint arXiv:1604.07370* (2016).

[32] Stab, C. and Gurevych, I.: Recognizing Insufficiently Supported Arguments in Argumentative Essays (2016).

[33] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijmm, T. A., Hirst, G. and Stein, B.: Computational Argumentation Quality Assessment in Natural Language, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 176–187 (online), available from ⟨http://www.aclweb.org/anthology/E17-1017⟩ (2017).

[34] Walton, D.: Arguing from Definition to Verbal Classification: The Case of Redefining 'Planet' to Exclude Pluto, *Informal Logic*, Vol. 28, No. 2, pp. 129–154 (2008).

[35] Walton, D., Reed, C. and Macagno, F.: *Argumentation Schemes*, Cambridge University Press (2008).