Towards Understanding Implicit Reasoning in Arguments via Multiple Warrants

Keshav Singh[†] Paul Reisert[‡,†] Naoya Inoue[*] Kentaro Inui[†,‡]

[†]Tohoku University, Sendai, Japan [*]Stony Brook University
[‡]RIKEN Center for Advanced Intelligence Project
keshav.singh29@ecei.tohoku.ac.jp paul.reisert@riken.jp
naoya.inoue.lab@gmail.com inui@tohoku.ac.jp

1 Introduction

In argumentative discourse such as debates and essays, some parts of the reasoning are unstated by the debater or writer, often for strategic reasons/purpose. Such implicit reasoning, hereby referred to as a *warrant*, is often inferred by listeners in distinct ways [10]. For example, consider the following argument consisting of a claim (i.e., declarative statement) and a premise (i.e., supporting statement):

(1) **Claim:** *Voting should be made compulsory*
Premise: *because it increases voter turnout.*

In Example 1, a potential set of warrants bridging the reasoning between the claim and the premise could then be considered as follows:

(2) **Warrant 1:***High voter turnout is good for a fair representation of society.*
Warrant 2:*Minorities who vote often have their needs prioritized over the majority.*

In the educational domain, practicing identification of such warrants has long been shown to improve one's argument comprehension skills [9, 4], thus aiding one to make better arguments [12]. For example Hillocks [8] helped students practice argumentation by making them write the appropriate warrant for a given argument which was later corrected if a better warrant existed. This helped students improve their critical thinking and writing skills.

Although identification and correction of warrants is an interesting challenge, especially when deployed in classrooms, despite the evidence, previous works with the aim to automate such identification of warrants have focused on methods which capture these implicit components from only a single perspective. Specifically, these methods usually focus on finding a single warrant for an argument when other possibilities exist. Habernal et al. [7] crowdsourced a pair of both correct and incorrect warrants per argument so that neural models would be able to distinguish between a good and bad warrant. Becker et al. [2] proposed an iterative process of warrant generation by experts aimed to create a single warrant unanimously decided by experts. However, previous work has mentioned that it is unrealistic to restrict on having one single warrant per argument due to vast number of ways to interpret the warrant [5, 10]. Hence, while the previous methods may lead to finding a single appropriate warrant, they do so at the expense of neglecting the other possible warrants.

In this work, we aim to collect multiple warrants per argument for a variety of topics through a systematic crowdsourcing process. In total, we collect warrants for 3 topics, which totals 6,200 warrants for 620 different arguments consisting of a claim and premise pair. Using a subset of the collected warrants for each topic, we analyze their overall quality through manual analysis and obtain a reasonable quality (Krippendorff's α=0.63). Furthermore, we conduct a preliminary analysis on the properties of warrants to better understand their structure.

2 Towards collecting multiple warrants

For collecting multiple warrants which can be potentially useful in real-world applications such as argument analysis or warrant correction, we require a dataset that fulfills the following criteria: i) diverse set of premises per topic to cater to maximum possible arguments, and ii) multiple unique warrants per premise. In addition, the dataset creation process must be cost effective without compromising data quality. Ideally, a dataset with multiple topics is desirable; however, collecting warrants across multiple topics is both difficult and time consuming. Thus, we define a simple metric to filter a handful of topics (specifically, 3) found in a large, well-known argumentation dataset of diverse premises in order to

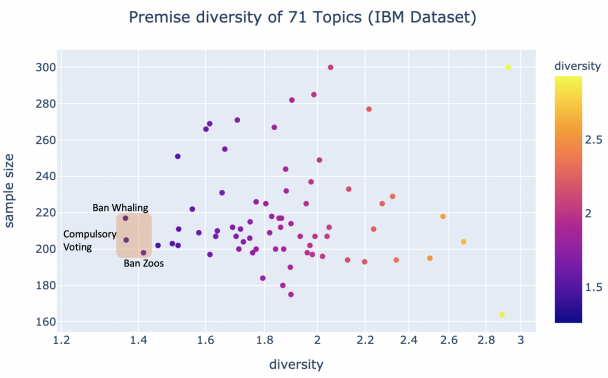

Figure 1: Scatter plot of premise diversity vs no. of premises for each topic (represented by points) in IBMRank30k corpus. Diversity of premises for each topic increases along the x-axis. For our task, we choose three least diverse topics for warrant crowdsourcing task.

start with a small and relatively easier topics, We describe our methodology for topic selection and collecting multiple user-generated warrants for a claim and premise pair in the following section.

2.1 Source data

As a source dataset, we opted for the arguments from *IBM-Rank-30K* dataset [6], which contains supporting and opposing stance arguments on 71 common controversial topics.

Several factors motivated our choice for *IBM-Rank-30K*: (1) Arguments were collected actively from crowdworkers with strict quality control measures as opposed to being extracted from targeted audiences such as debate portals. (2) The arguments have already been annotated with point-wise quality. These factors provide a vast majority of all the possible arguments that can be made for a given topic with quality scores.

To select the topics for our warrant crowdsourcing task, we define a new metric. Specifically, for each topic(t), we estimate the diversity of premises(d_t) as a function of its vocabulary size:

$$d_t = \frac{(|V(p_t)| - |V(p_t^{50\%})|)}{(|p_t| - |p_t^{50\%}|)}, \qquad (1)$$

where p_t is a set of premises associated with t, $p_t^{50\%}$ is a 50% random sample of p_t, and $V(p)$ is a set of unique words (i.e. vocabulary size) of p obtained after tokenization and lemmatization.

The final diversity (d_t) for each topic is calculated after averaging over multiple random resampling runs. As shown in Fig 1, the variety of premises for topics in *IBM-Rank-30k* is strongly dependent on

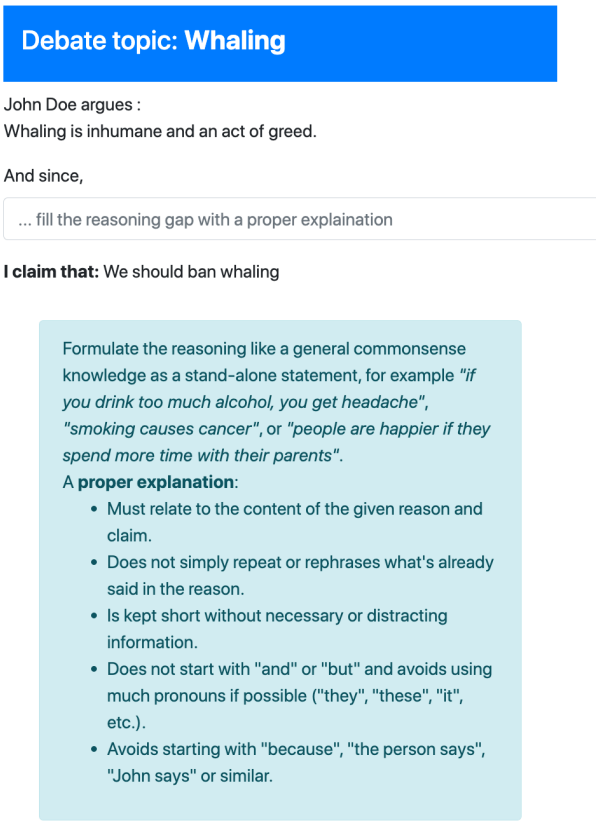

Figure 2: Crowdsourcing interface for collecting warrants for a given topic and premise. Workers were additionally advised to adhere to simple rules when writing their warrants.

the topic. To create our warrant KB, we choose 3 topics with the least growth rate; namely i) Whaling should be banned (d_t : 1.36), ii) Voting should be made compulsory (d_t : 1.37), and iii) Zoos should be banned (d_t : 1.41).

2.2 Crowdsourcing methodology

For collecting warrants for arguments across the 3 pre-selected topics, we conducted a crowdsourcing task using Amazon Mechanical Turk [1] (AMT) platform. Prior to deploying our task in full, we conducted a preliminary study for creating our guidelines by annotating a subset of warrants crowdsourced via multiple trial runs to understand both good and bad warrants. As shown in Fig 2, we provide crowdworkers with an argument (i.e., claim and a premise) and instruct workers to provide a warrant that fills the implicit reasoning between the claim and premise in the form of natural sentence. Additionally, for maintaining high quality annota-

[1] www.mturk.com

Topic *T*: *We should ban whaling*
Premise *P*: *Alternatives are available to nearly every product produced from whales.*

Warrants *W (Score: 2)* ->
- We don't need the products that result from whaling
- There are different fish that can be alternative for whale meat

Warrants *W (Score: 0)* ->
- Whales are an endangered and are vital to the health of the marine environment
- Whales are animals vital for the balance of the ocean[1]

Topic *T*: *We should introduce compulsory voting*
Premise *P*: *Compulsory voting would increase voter turnout.*

Warrants *W (Score: 2)* ->
- A greater voter turnout leads to a more representative government
- Increase in voter turnout will eradicate the politicians that are dependent on few individuals which can control the election

Warrants *W (Score: 0)* ->
- It is not democracy if there is only 50 percent of voter turnout. higher voter turnout makes government more representative
- People will complain about politicians regardless of voter turnout.

Figure 3: Examples of warrants captured by our crowdsourcing method. Warrants acquired from different workers as scored best by experts are shown on the left and the worst scored are shown on the right. Note that for a few warrant instances, we found very similar warrants after analyzing them into individual clausal sentence as shown in red.

tions, we filtered crowdworkers via our custom Reasoning Qualification Test (RQT). In RQT, workers were asked 3 simple reasoning questions aimed to test their ability to identify warrants. Only those who answered all questions correctly were allowed to proceed to the final annotation task.

3 Collected warrant statistics

In this section, we describe our current set of warrants collected through crowdsourcing. To assess the quality, we also create guidelines for experts to manually examine the crowdsourced warrants.

In total, we collect 6,200 warrants in total for 620 unique claim and premise pairs (10 warrants per premise) from 177 unique crowdworkers. The specific amount of warrants collected per each topic are as follows; Abolish zoos: 198, Introduce compulsory voting: 205 and Ban whaling: 217.

3.1 Quality

To examine the quality of the crowdsourced warrants, we ask 2 annotators, both experts of argumentation and authors of this paper, to judge the quality of 200 randomly sampled warrants on a 0-2 scale (0:weak warrant, 2:strong warrant). As shown in Table 1, the Krippendorff's α (interval) and Cohen's Kappa between both annotators for 200 randomly sampled set was 0.63 and 0.42 respectively. These scores correspond to moderate agreement and

Topic: We should	α	κ	A1	A2
Abolish zoos	0.64	0.42	1.60	1.50
Introduce compulsory voting	0.53	0.39	1.46	1.55
Ban whaling	0.65	0.43	1.49	1.25
Overall	0.63	0.42	1.51	1.43

Table 1: Topic-wise quality statistics for the collected warrants. α and κ are the Krippendorff's and Cohen's inter-annotator agreement scores respectively. A1 and A2 are the average scores given to the warrants by our two expert annotators.

are comparable to agreement levels in similar computational argumentation works [1, 3]. Additionally, as shown in Fig 3, the collected warrants ranged from being weak to strong, where some being totally unrelated to the argument (scored:0) while majority explicating the reasoning (scored:2).

3.2 Warrant properties

We perform a manual check on a subset of 2,000 warrants from 200 arguments (i.e., 10 warrants per argument) to determine if the collected warrants are *related* to the argument. On an average, we discovered that 6.67/10 (66 .7%) warrants were related while the rest were either nonsensical or unrelated. During our analysis, we also found that many warrants were compositional statements i.e. comprising

Figure 4: Example of reasoning patterns on top of collected warrants. Both warrants help infer the connection between the claim and the premise more clearly.

of two or more compound statements. We observed that many of the warrants contained repetitive statements, as shown in Fig 3.

Towards understanding the underlying reasoning, we conducted a preliminary analysis on warrants for determining what properties of them differ for a claim and premise pair. Motivated by a set of reasoning patterns [11] used to represent various schemes in argumentation [13], we analyze roughly 48 warrants for 6 claim-premise pairs (8 warrants per pair) to determine the coverage of warrants which can be represented with reasoning patterns. We find that roughly 20/48 (41.67%) warrants can be decomposed into reasoning patterns. An example of such annotation is shown in Fig 4. In this example, the reasoning that *whaling* suppresses *animals* and *whaling* suppresses *rights* is implicit, which the warrants help explicate. Given the coverage of reasoning patterns in our analysis, we will consider ways to improve our warrant collection task by utilizing them in our future work.

4 Conclusion and future work

In this work, we tackle the difficult task of collecting warrants in arguments. We developed a methodology for collecting multiple warrants for arguments and utilize crowdsourcing for collecting 6,200 warrants for 620 arguments. For testing the integrity of the warrants, we conduct a small annotation study on 200 warrants and obtain a reasonable annotator agreement (0.63 Krippendorff's α). Finally, we conduct a preliminary analysis on the decomposition of the warrants and find on average 3 warrants collected per argument to comprise of more than one warrant.

In our future work, we will expand our dataset for a variety of different topics and devise an approach for decomposing collected warrants to a single clausal form. We will also test the usefulness of our collected warrant knowledgebase for the task of warrant explication for unseen arguments in a realistic setting, such as deploying our warrant knowledgebase in schools in which students debate and/or write essays on controversial topics.

References

[1] Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. From arguments to key points: Towards automatic argument summarization. *arXiv preprint arXiv:2005.01619*, 2020.

[2] Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. Enriching argumentative texts with implicit knowledge. In *International Conference on Applications of Natural Language to Information Systems*, pages 84–96. Springer, 2017.

[3] Filip Boltužić and Jan Šnajder. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Sibel Erduran, Shirley Simon, and Jonathan Osborne. Tapping into argumentation: Developments in the application of toulmin's argument pattern for studying science discourse. *Science education*, 88(6):915–933, 2004.

[5] James B Freeman. Relevance, warrants, backing, inductive support. *Argumentation*, 6(2):219–275, 1992.

[6] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*, 2019.

[7] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] George Hillocks. Ej in focus: Teaching argument for critical thinking and writing: An introduction. *The English Journal*, 99(6):24–32, 2010.

[9] David Hitchcock and Bart Verheij. *Arguing on the Toulmin model*, volume 10. Springer, 2006.

[10] Christian Kock. Multiple warrants in practical reasoning. In *Arguing on the Toulmin model*, pages 247–259. Springer, 2006.

[11] Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[12] Sarah von der Mühlen, Tobias Richter, Sebastian Schmid, and Kirsten Berthold. How to improve argumentation comprehension in university students: Experimental test of a training approach. *Instructional Science*, 47(2):215–237, 2019.

[13] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, 2008.