Exploring Methodologies for Collecting High-Quality Implicit Reasoning in Arguments

Keshav Singh[‡] **Farjana Sultana Mim**[‡] **Naoya Inoue**[⋆,†]
Shoichi Naito[‡,†,◇] **Kentaro Inui**[‡,†]

[‡] Tohoku University [†] RIKEN [⋆] Stony Brook University [◇] Ricoh Company, Ltd.
{keshav.singh29,naoya.inoue.lab}@gmail.com
{mim.farjana.sultana.t3,naito.shoichi.t1}@dc.tohoku.ac.jp
inui@tohoku.ac.jp

Abstract

Annotation of implicit reasoning (i.e., warrant) in arguments is a critical resource to train models in gaining deeper understanding and correct interpretation of arguments. However, warrants are usually annotated in unstructured form, having no restriction on their lexical structure which sometimes makes it difficult to interpret how warrants relate to any of the information given in claim and premise. Moreover, assessing and determining better warrants from the large variety of reasoning patterns of unstructured warrants becomes a formidable task. Therefore, in order to annotate warrants in a more interpretative and restrictive way, we propose two methodologies to annotate warrants in a semi-structured form. To the best of our knowledge, we are the first to show how such semi-structured warrants can be annotated on a large scale via crowdsourcing. We demonstrate through extensive quality evaluation that our methodologies enable collecting better quality warrants in comparison to unstructured annotations. To further facilitate research towards the task of explicating warrants in arguments, we release our materials publicly (i.e., crowdsourcing guidelines and collected warrants).

1 Introduction

Implicit reasonings, commonly referred to as *warrants* (Toulmin, 1958), have long been studied to understand the grounds on which a premise lends support to the claim (Freeman, 1992). In other words, a warrant, when made explicit, clearly shows the inferential link between claim and premise (Pineau, 2013). As depicted in Figure 1, identification of such warrants by students has been shown to aid them in making better arguments (Erduran et al., 2004), as well as improving their critical thinking skills (von der Mühlen et al., 2019) and argument comprehension process (Hitchcock and Verheij, 2006).

While explication of warrants with assistance from teachers has been shown to be useful for improving students' argumentation skills, automating this explication process would not only help students to be less dependent on teachers, but it can also be beneficial for different downstream educational applications such as argument analysis (Becker et al., 2020), enthymeme reconstruction (Razuvayevskaya and Teufel, 2017; Hulpus et al., 2019) and essay scoring (Williamson, 2013). However, building an automated warrant explication system has been a challenge due to the difficulty of collecting warrants in a form that explicitly manifests the way a warrant relates the information between claim and premise. Generally, warrants are annotated in an unstructured (i.e., free-text) format which lays no restriction on its lexical structure (Boltužić and Šnajder, 2016). As a result, sometimes the warrant consist of no information that overlaps with claim or premise which makes it difficult to understand how the warrant connects the claim to its premise. Furthermore, for a given argument, unstructured warrants can be framed in diverse ways that would have a wide variety of reasoning patterns (Kock, 2006) (given no restriction on the lexical structure), and identifying the correct ones from this large pool of warrants can be a preposterous task.

In order to annotate warrants on a large-scale and in a way that overcomes the aforementioned challenges, we propose two novel warrant annotation methodologies: Pre-defined Keyword-based Warrant (PKW) and User-defined Keyword-based Warrant (UKW), which restrict a warrant's lexical structure to a semi-structured form. In contrast to approaches that crowdsource unstructured warrants, these methodologies explicate warrant by enforcing it to have the key information (i.e., keywords) from both claim and premise. The intuition behind our semi-structured approach is to restrict the structure of warrants to specific keyword-based

Proceedings of The 8th Workshop on Argument Mining, pages 57–66
Punta Cana, Dominican Republic, November 10–11, 2021.

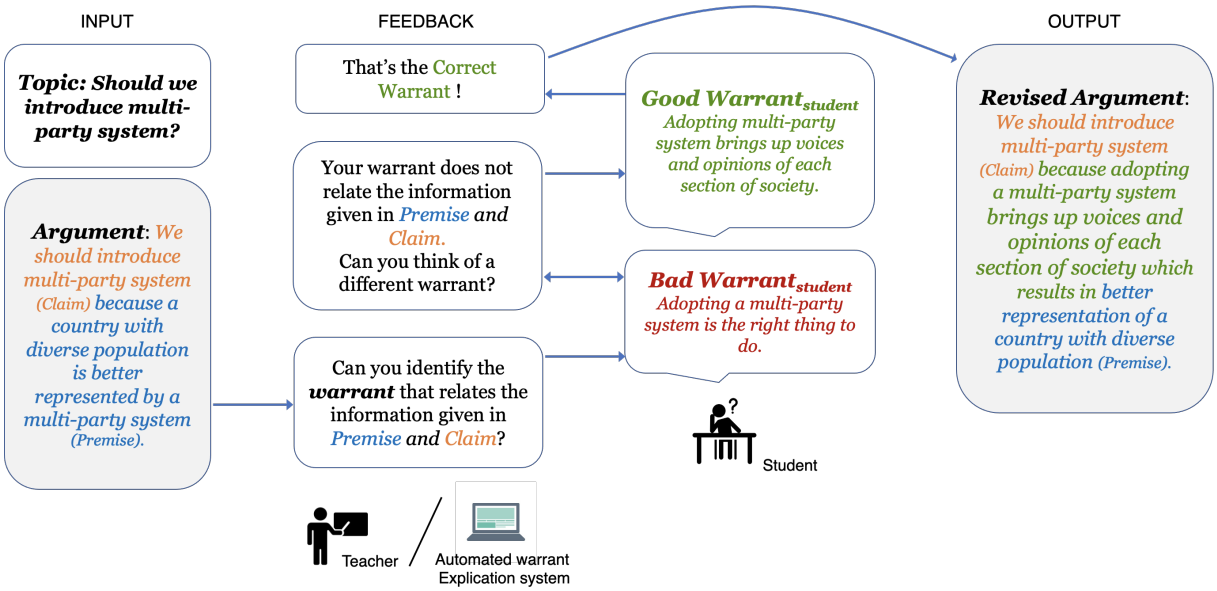

Figure 1: A typical example of warrant explication. Feedback provided by teachers or automated warrant explication system can help students identify correct warrants and leverage it in their revised argument to make the argument more reasonable.

reasoning patterns. We hypothesize that leveraging such keyword-based pattern could assist in collecting high quality warrants, where keywords are derived from the original argument (i.e., claim and premise). Our assumption follows the formal definition of warrants in the sense that warrants act as a inferential link between the contents of claim and its premise (Toulmin, 1958; Freeman, 1992).

Many previous work demonstrated different strategies of collecting high quality warrants (Becker et al., 2017; Habernal et al., 2018; Becker et al., 2020), but did not apply any restriction on the structure of warrants to handle the variety of reasoning patterns in which warrants can be explicated. In contrast, our annotation methodologies are designed to restrict the reasoning pattern and ensure that warrant explicates the reasoning link between claim and premise.

In order to evaluate the warrants annotated via our proposed UKW and PKW methodologies, we devise specific guidelines to judge their quality through scoring. We also collect unstructured warrants (i.e., Natural Language Warrants (NLW)) and perform quality evaluation on them in order to compare with our annotated warrants. Our results suggest that in comparison to NLW, high quality warrants can be annotated via our proposed UKW methodology. To the best of our knowledge, this is the first study which targets large scale annotation of warrants in semi-structured form. To facilitate further research in warrant explication, we publish our crowdsourcing guidelines and the preliminary corpus of around 1700 warrants that are annotated via UKW methodology, covering over 600 arguments [1].

2 Related Work

Explication of warrants in arguments has already been approached in many previous researches. In an initial attempt, Feng and Hirst (2011) proposed leveraging argumentation schemes (Walton et al., 2008) as a means to automatically reconstruct warrant, but did not approach the task due to the absence of training datasets. To overcome the unavailability of dataset, Boltužić and Šnajder (2016) leverage crowdsourcing and ask non-expert workers to annotate all possible variations of warrants for a given claim-premise pair. However, they concluded that the annotation of warrants varied both in number annotated per argument and in content due to no restrictions imposed on in the annotation process.

In order to overcome the prior difficulties, recent approaches leverage crowdsourcing to restrict the number of warrants collected per argument and either employ a step by step filtering process to weed out bad warrants (Habernal et al., 2018) or hire ex-

[1]Our crowdsourcing guidelines and annotated warrants are publicly available at https://github.com/cl-tohoku/ukw-warrants

	Argument	Warrant
1.	Claim: We should abolish zoos. Premise: Zoos are notorious for animal abuse.	Abolishing zoos *leads to animals being in their natural habitat which results in no* animal abuse.
2.	Claim: We should ban whaling. Premise: Whaling is considered to be unacceptable cruelty towards animals.	Banning whaling *would stop the inhumane methods of stabbing whales which is* unacceptable cruelty towards animals.
3.	Claim: We should introduce compulsory voting. Premise: Compulsory voting can help obtain better results during elections.	Introducing compulsory voting *leads to every citizen exercising the right to vote which can help* obtain better results during elections.

Table 1: Example warrants demonstrating the semi-structured form used for annotation.

perts who iteratively converge to a single warrant annotation (Becker et al., 2017). In an advanced attempt, Razuvayevskaya and Teufel (2017) explored whether it is feasible for human annotators to explicate warrants in arguments. They propose the idea of template-based warrant reconstruction using information from premise and claim, and employ experts to perform the task. In contrast, our annotation methodology is designed for expert as well as non-expert workers. Additionally, our annotation aims to restrict the warrant's structure to semi-structured format such that it comprises knowledge that is necessary to form an inferential link between the key contents of claim and premise. Furthermore, we do not restrict the number of warrants to one, but collect set of warrants per argument that fulfill the criteria of qualifying as a high quality warrant.

3 Warrant Desiderata

We define warrant that we want to annotate by characterizing its desired properties. Properties such as structure of warrant, quality in terms of how well a warrant links claim and premise, and feasibility of annotating warrant for an argument. In this work, we define this feasibility in terms of whether a warrant can be annotated for an argument, regardless of whether it is good or bad. We assume that explication of a warrant might not be possible if the argument is too good (i.e., warrant is explicated in the premise) or if the argument is too bad (i.e., no warrant can explicate the link between claim and premise).

Structure Warrants are implicit reasoning that logically link the contents of claim and premise (Toulmin, 1958; Freeman, 1992). Therefore, we hypothesize that a warrant should have a structure that (a) comprises the key information given in claim and premise, and (b) explicates logical connection between the aforementioned key information with some implicit knowledge that is relevant to the argument. We define the warrants framed in such a way as semi-structured warrants. Examples of such semi-structured warrants are provided in Table 1, where key information or keywords from claim and premise is linked with relevant implicit knowledge which all together forms a semi-structured warrant.

Feasibility Warrants may not be explicable for all the arguments. Specifically, for an argument with bad premise there may be no feasible way to explicate the logical link between claim and premise. For example, the warrant for the argument *"We should introduce multi-party system because it's the right thing to do"* is not feasible, since the argument is a fallacy (i.e. begging the question) where the premise: *"it's the right thing to do"* provides no adequate support to the claim: *"We should introduce multi-party system"*. Similarly, for arguments with very good premise, it might not be necessary to explicate the warrant since the warrant might already be explicated in the premise. In contrast, as shown in Table 1, for arguments with moderately good/bad premise, we assume that one can frame a warrant by leveraging argument relevant external knowledge. [2].

[2]The arguments shown in Table 1 were already annotated with moderate quality scores in a larger study (Gretz et al., 2019)

Quality A key factor distinguishing warrants from any other type of implicit knowledge or statement (e.g., commonsense knowledge) is their ability to justify the flow of reasoning between claim and premise. For example, in Table 1, warrant (3) explicitly answers how introducing compulsory voting can help obtain better results in elections. These are the type of warrants we would like to annotate. Conversely, statements which do not serve this purpose cannot be qualified as a warrant. For example, given argument (3) from Table 1, the statement *"Introducing compulsory voting enables people to freely choose their favourite candidate which results in encouraging better results during elections"* cannot be labeled as a warrant because it offers little to no help in bridging the implicit reasoning link between claim and premise.

4 Annotation Methodologies

In this section, we discuss the development and design of our proposed semi-structured annotation methodologies. In particular, we consider two methodologies for annotating warrants: *Pre-defined Keyword-based Warrant* and *User-defined Keyword-based Warrants*.

Pre-defined Keyword-based Warrants (PKW) In order to annotate semi-structured warrants that encompass information from claim and premise, we propose using keywords which encode the key information given in claim and premise. As shown in Table 1, the purpose of these keywords is to create a semi-structured format for completing a warrant annotation such that keywords from claim form the initial (shown in red) and keywords from premise (shown in blue) form the latter part of the warrant. The keywords are linked by implicit knowledge that is necessary to connect the keywords in a meaningful way. Example annotation and task design of PKW annotation is shown in Figure 2. For PKW methodology, the annotator is initially provided with keywords and is tasked to explain the flow of reasoning between them by writing implicit knowledge (i.e., hidden reasoning).

In order to provide pre-defining keywords to the annotator, we employ spaCy (Honnibal et al., 2020) and automatically extract the key information from claim and premise by parsing the sentence into verb/noun phrases. For example, for claim *"We should ban whaling"*, the verb phrase *"Banning whaling"* and for premise *"Whales are necessary for ecological sustainability of the oceans"*, the noun phrase *"ecological sustainability of the oceans"* is extracted. To ensure the feasibility of framing a warrant with extracted keywords, we perform a manual check and if needed, make minimal changes to its tense or word-order.

User-defined Keyword-based Warrants (UKW) While PKW methodology introduces restrictions on warrant annotation via pre-defined keywords, they might be too restrictive or not provide sufficient flexibility for annotators to annotate the warrant. Moreover, automatically extracting keywords from claim and premise can be sometimes challenging due to varied syntactic structure of the argument. Therefore as an alternate approach, we ask the annotators to derive their own keywords from claim and premise. To do this, we provide detailed guidelines and concrete examples in our interface for annotators so that they can correctly understand the process of deriving keywords.

The annotation design and example annotation is shown in Figure 3 , where for the given claim and premise, the keywords from claim can be a verb phrase: *"Introducing compulsory voting"*, and keywords from premise can be verb phrase: *"obtain better results during elections"*. To avoid annotations where annotator might write keywords with information from outside claim/premise, annotators were strictly advised not to use any external knowledge when writing the keywords, although the use of external knowledge was permitted for writing the hidden reasoning.

5 Warrant Collection Procedure

Our goal is to establish a procedure for collecting semi-structured warrants and their annotations at large-scale. In order to collect such warrants from each methodology, we build a multi-step crowdsourcing process designed for encouraging annotator's creativity, while preventing biases in the annotations.

In general, we break the warrant collection procedure into three steps of simple tasks: (i) Deriving keywords, (ii) Judging feasibility and (iii) Framing warrant. In addition, we implement several mechanisms for quality assurance and employ manual checks to ensure annotators understand and perform the final task correctly.

(i) Deriving keywords In order to collect semi-structured warrants, we require keywords derived from each of claim and premise. These keywords

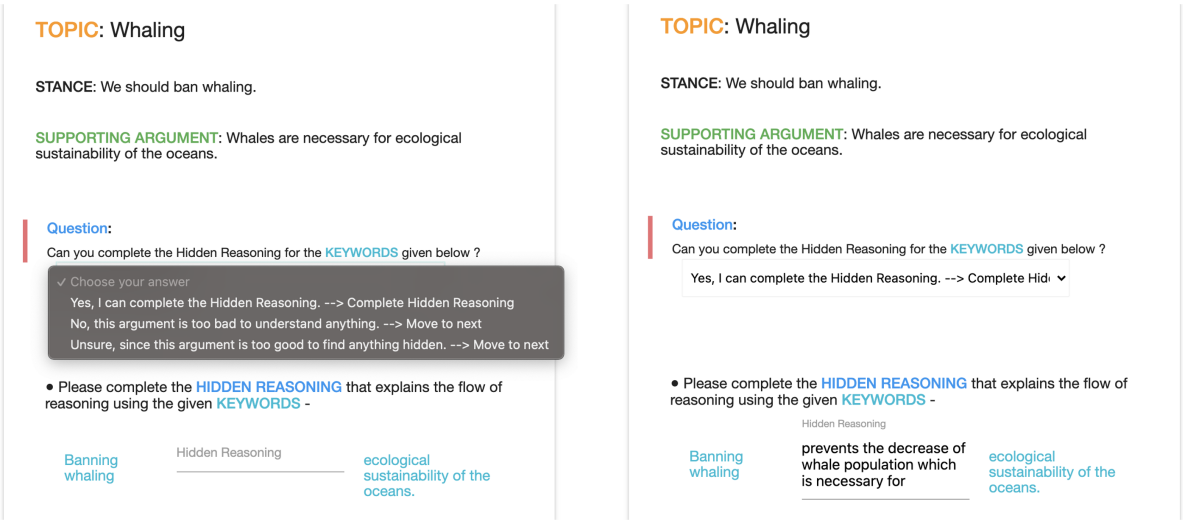

Figure 2: The interface used for warrant annotations, along with an example of a annotator's annotation for Pre-defined Keyword-based Warrant. To avoid using complicated terminology, we used the terms *stance*, *supporting statement* and *hidden reasoning* to convey the notion of claim, premise and warrant respectively.

act as the skeleton of the final annotated warrant. For example, as shown on the left side of Figure 3, *"Introducing compulsory voting"* and *"obtain better results during elections"* are keywords derived from claim and premise respectively. To derive these keywords, annotators are instructed to strictly include only key information conveyed in their respective counterparts and no external information. For PKW methodology, keywords in this step are already derived as shown in Figure 2. To ensure annotators understand the notion of keywords, we provide sufficient variety of examples for them to get used to this sub-task.

(b) Judging feasibility Since warrants may not be explicable for all the arguments, i.e., if a premise is very bad or contrastingly very good, we explicitly ask annotators to judge the feasibility of writing a warrant by asking if they can complete the hidden reasoning. This step is rather tricky since annotators may be biased to answer "No" or "Unsure" (See Question in Figure 2 and 3) to avoid doing the task and finish the task quickly. To avoid this, we treat this step as bonus question and depending on majority response i.e., if majority of annotators believe a warrant can be explicated for the given argument, then the majority annotators get bonus. Similarly, if majority of annotators believe a warrant cannot be explicated for a given argument, then again majority annotators get bonus. We keep a high bonus for this step in order to compel annotators to do task as instructed instead of providing low quality response. This step helps us get a better judgement of feasibility of warrants and also identify annotators who are not doing the task properly.

(c) Framing Warrant The last step in warrant collection procedure is for the annotators to frame the hidden reasoning to complete the warrant. This step is the most challenging since it requires the annotator to be logical and use his background knowledge to complete warrant annotation. To complete this via PKW methodology, annotator's are restricted in terms of pre-defined keywords, while for UKW methodology the annotator can annotate the warrant by minimally changing the keywords as well as hidden reasoning.

Auxiliary verification measures For each task, we hold preliminary qualification test that consists of several basic questions to judge the understanding of annotator's reasoning skills. Annotators who score more than a pre-defined threshold ($\geq$ 80%) are granted access to the main task. Our qualifications are open to annotators from major English speaking countries, namely USA, UK, New Zealand and Canada. Additionally, to address any ethical issues (Adda et al., 2011) raised by our task, we actively monitored multiple pilot tests to ensure annotators were satisfied with our task. Simultaneously, we corresponded directly to annotators that had questions/comments on our task. Annotators were paid in accordance with the minimum wage

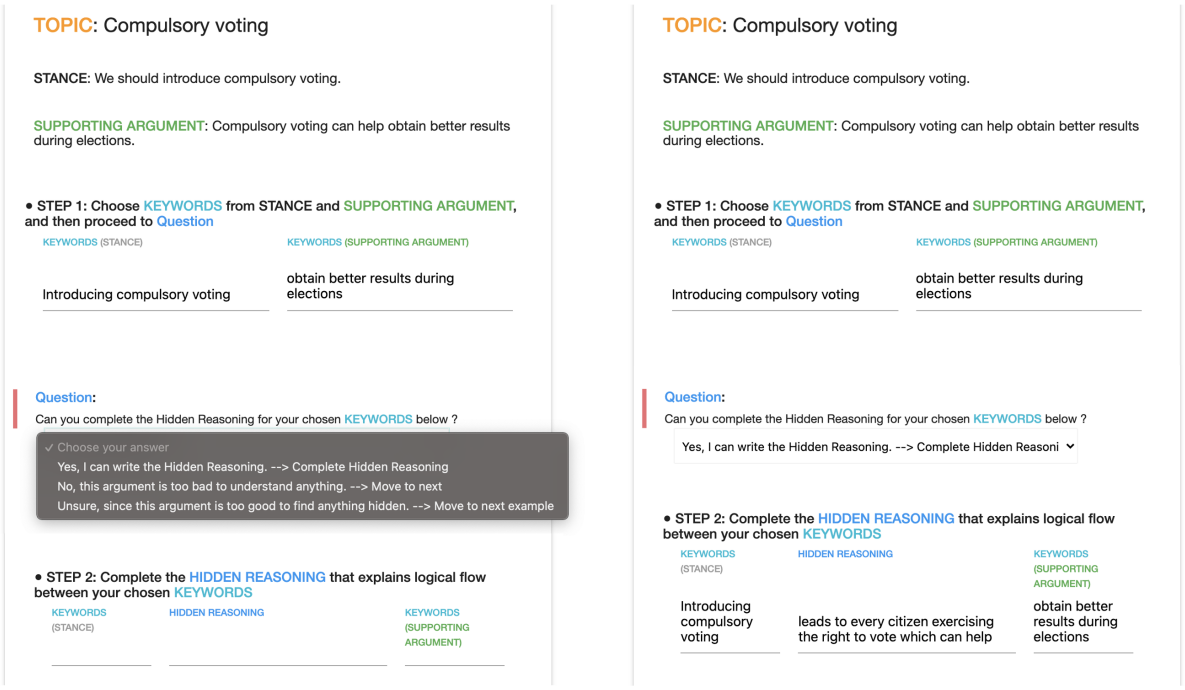

Figure 3: The interface used for our warrant annotations, along with an example of a annotator's annotation for User-defined Keyword-based Warrant.

calculated by conducting many trials and based on average work-time. Additional bonus was paid to annotators that provided quality annotations.

6 Crowdsourcing

We choose Amazon Mechanical Turk (AMT)[3] as our crowdsourcing platform due to its success in previous argumentation mining tasks (Habernal et al., 2018). As an initial step, we only allowed annotators who had $\geq$ 98% acceptance rate and $\geq$ 5,000 approved Human Intelligence Tasks (HITs). Each annotator was paid \$0.30 for doing the task and workers who qualified for bonus were paid an additional \$0.25 per HIT. For the task involving UKW annotations, the payment was increased to \$0.40 for doing the task and \$0.35 as bonus.

Source data For the purpose of annotating warrants for a given set of arguments, we utilize a well-known argumentation dataset, IBM-Rank-30K corpus (Gretz et al., 2019), which already consists arguments in the form of claim and premise. The dataset contains around 15K crowd-sourced arguments covering 71 topics, annotated for supporting as well as opposing stance.[4] The arguments were collected with strict length limitations and accompanied by extensive quality control measures. Our inspection of arguments from IBM-Rank-30k revealed that a for large proportion of the arguments we can explicate warrants. To proceed with our warrant annotation, we selected a subset of three 3 well-known debatable topics: *We should abolish zoos*, *We should ban whaling* and *We should introduce compulsory voting*.

Annotations In addition to warrants annotated via PKW and UKW methodologies, we also annotated unstructured warrants via crowdsourcing to compare the quality of annotated warrants across different methodologies. We followed similar crowdsourcing procedure as employed by Habernal et al. (2018) and refer to the warrants annotated via this process as *Natural language Warrants (NLW)*. For each methodology, we annotate 40 unique claim and premise pairs, randomly chosen from the three pre-selected topics. Each argument was annotated by 5 annotators resulting in a total of 200 annotations per methodology and all warrants were limited to have a length between 60 and 200 characters excluding keywords.

Filtering For each methodology, it is possible to collect at most 200 warrants (given 40 arguments and 5 annotators). However, not all annotators

[3] www.mturk.com

[4] In our work we only focus on arguments with supporting stance.

Methodology	Natural Language		User-defined		Pre-defined	
Topic	α	Avg.	α	Avg.	α	Avg.
Abolish zoos	0.64	1.55	0.67	1.60	0.62	1.57
Intro. compulsory voting	0.45	1.50	0.51	1.55	0.53	1.47
Ban whaling	0.63	1.37	0.61	1.61	0.58	1.59
Overall	0.57	1.46	0.56	**1.59**	0.53	1.55

Table 2: Comparison between the different warrant crowdsourcing methodologies. Avg. corresponds to the average score given by both expert annotators.

Score	Explanation
0	Warrant is unrelated to the claim and its premise.
1	Warrant is related to the claim and premise but does not make the relationship between them easy to understand and/or strengthen the argument. In addition, the warrant may overlap or be a paraphrase of the premise.
2	The relationship between the claim and premise is easier to understand and/or strengthened because of the warrant.

Table 3: Guidelines used by our expert annotators for scoring the quality of warrants on a scale of 0-2.

chose "Yes" when they were asked to judge if they can write a warrant. After filtering the negative responses at this step, we discover that, in total, annotators wrote 155, 101 and 65 warrants out of possible 200 warrants for Natural language Warrant, PKW and UKW methodologies respectively. We utilize these 321 collected warrants for further analysis.

7 Results

In order to analyze the quality of the annotated warrants, we hired two annotators who are experts in the field of argumentation to score the crowdsourced warrants. To frame quality scoring guidelines for judging warrants, we ran several pilot tests and take expert advice to make our guidelines easier to interpret. Our final quality annotation guidelines are shown in Table 3. The experts were asked to score a given warrant on a scale of (0-2), with 0 being the lowest and 2 being the highest. Both annotators were given 50 warrants randomly chosen from the pool of collected warrants for each methodology. Each topic was represented fairly in the quality annotation step with each topic having at least 15 warrants. In total, our experts annotated 150 warrants out of a total 321 annotated warrants. As shown in Table 2, the agreement between both experts as judged by Krippendorff's alpha α (Krippendorff, 2011) was found to be fairly good. We find that the average scores given by two expert annotators (Avg.) on a scale of (0-2) indicated user-defined and pre-defined methodologies with overall higher quality warrants as compared to natural language warrants. We also measure the combined average score given to the warrants for each topic to measure if the warrants belonging to one topic was of higher quality. We find that on average warrants annotated for *Introducing compulsory voting* were scored the lowest while for other topics was fairly high.

7.1 Qualitative Analysis

To further analyze the quality of warrants and the quality of the entire crowdsourcing process, we analyze sample of the warrants collected via each methodology and which were annotated by experts.

In Table 4, we can see that the warrants that were scored the highest by experts have fair amount of keyword overlap with the claim/premise. This follows from our initial motivation for using semi-structured annotation methodology where we hypothesized that the inclusion of keyword information from the claim and premise can assist in framing better quality warrants. As shown in Table 4, warrants UKW (1) and PKW (1) have similar keywords *"Banning whaling"* and *"ecological sustainability of the oceans"*, hence the annotated implicit knowledge is also same. This indicates that our keyword-based PKW and UKW methodologies restrict the diverse ways in which warrants can be explicated. On the contrary, for NLW (1) the warrant is analogous to a general statement yet has

1.	Claim: We should ban whaling. Premise: Whales are necessary for ecological sustainability of the oceans.
NLW: *PKW*: *UKW*:	Marine life in the ocean cannot survive without ecological sustainability. Banning whaling would prevent the decreasing of whale population on which marine life thrives which will result in ecological sustainability of the oceans. Banning whaling prevents the decreasing of whale population which is necessary for ecological sustainability of the oceans.
2.	Claim: We should abolish zoos. Premise: We should abolish zoos to prevent the cruel confinement of wild animals.
NLW: *PKW*: *UKW*:	Animals in confinement suffer physically and emotionally. Abolishing zoos enables animals to live a more stimulating and fulfilling life in the wild which prevents the cruel confinement of wild animals. Zoos force many animals to live in prison-like environment with unhygienic conditions which is a cruel way of confining wild animals.
3.	Claim: We should introduce compulsory voting. Premise: Compulsory voting can help encourage better results during elections.
NLW: *PKW*: *UKW*:	Mandatory voting will reflect people's actual preferences in election results. Compulsory voting produces a winning candidate that is more accurately representative of all the voters which ensures better election results. Introducing compulsory voting stops one side from rigging the process by canvasing more people which results in better elections.

Table 4: Examples of warrants scored 2 by both experts. Majority of warrant belonging to UKW methodology were found to be of good quality.

1.	Claim: We should ban whaling. Premise: Whaling has led to a major decrease in whale populations over the years.
NLW: *PKW*: *UKW*:	Whales could soon die off completely. It is our duty to ban whaling. Banning whaling leads to whales not dying which has cause decrease in whale population over the years. Banning whaling is a harmful practice which results in decrease in whale population over the years.
2.	Claim: We should abolish zoos. Premise:It is unfair to trap animals from their natural habitat and confine them to small spaces for human entertainment.
NLW: *PKW*: *UKW*:	Zoos keep animals captive where where they cannot run free and thrive. Abolishing zoos is a bad practice because it is unfair to trap animals from their natural habitat and confine them to small spaces for human entertainment. Abolishing zoos makes sure that animals are not being treated unfair and in small spaces.
3.	Claim: We should introduce compulsory voting. Premise: Compulsive voting is a patriotic act that must be fully complied with.
NLW: *PKW*: *UKW*:	We enjoy the freedom and liberty enjoyed by expressing our opinions. We should take advantage of such compulsion. Introducing compulsory voting makes you feel that you belong to your country which is a patriotic act. Compulsory voting will force people to engage in politics and choose a right leader that can lead their country in the future.

Table 5: Examples of warrants scored 0 or 1 by both experts. Low scored warrants were mainly paraphrased (Scored 1) from the premise or did not relate to the given topic (Scored 0).

similar implicit knowledge explicated in UKW (1) and PKW (1) warrants. This hints that NLW can also be of higher quality but its quality can vary to a larger extent due to no restrictions.

In Table 5, we can see that even though the warrants encode claim-premise information, the quality of the warrant can essentially be bad. For example, the warrants PLW (1) and UKW (1) explicate implicit knowledge which does not make the inferential link between claim and premise clear. Additionally, we note that while keyword-based methods restrict most warrants to a single sentence, natural language warrants often consist of shorter yet multiple sentences. Overall, we found that 23% of natural language warrants were composed of more than one sentences. Besides, such warrants were scored low and were often found to be paraphrased from the information in the premise or annotators rephrased previous premise in place of the warrant. Although this was mostly observed in natural language warrants, we found few such instances in our keyword-based collected warrants (See UKW (1) in

Table 5). This suggests that while keyword-based methods assist in collecting warrants that explicate the inferential link between claim and premise, it still does not guarantee high quality warrant annotations and might require further adjustments.

7.2 Preliminary large-scale corpus

Based on our finding that user-defined keyword-based warrant methodology comparatively results in better warrants, we follow this method to collect a total of 1700 warrants across 3 topics, annotated for 600 claim-premise pairs. All warrants are limited to have a length between 60 and 200 characters. Since this is an ongoing work, we plan to make further analysis on our preliminary dataset in future.

8 Conclusion and Future work

In this work, we tackle the difficult task of explicating warrants in arguments and propose two novel methodologies to annotate warrants in semi-structured format. We conduct extensive analysis and perform annotation study for determining the appropriate methodology for collecting warrants and show that user-defined keyword based approach produces the highest quality warrants as compared to pre-defined keyword-based warrant and natural language warrants. In future, we plan to extend the annotation of warrants for more diverse topics. Moreover, we plan to cover warrant annotations for claim-premise pairs with premises attacking the claim in addition to premises supporting the original claim. We would also like to test the usefulness of our annotations for constructing a model for automatic warrant explication which can be used to explicate warrant for any given argument. We believe that such a model can be useful in a pedagogical setting to perform downstream tasks such as argument analysis or giving constructive feedback to students.

Acknowledgements

This work was partially supported by JST CREST Grant Number JPMJCR20D2 and NEDO Grant Number J200001946. The authors would like to thank Paul Reisert, other members of the Tohoku NLP Lab, and the anonymous reviewers for their insightful feedback.

References

Gilles Adda, Benoît Sagot, Karën Fort, and Joseph Mariani. 2011. Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use. In *5th Language and Technology Conference*.

Maria Becker, Ioana Hulpuş, Juri Opitz, Debjit Paul, Jonathan Kobbe, Heiner Stuckenschmidt, and Anette Frank. 2020. Explaining arguments with background knowledge. *Datenbank-Spektrum*, 20(2):131–141.

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

Filip Boltužić and Jan Šnajder. 2016. 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. Association for Computational Linguistics.

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

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

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

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

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

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ioana Hulpus, Jonathan Kobbe, Christian Meilicke, Heiner Stuckenschmidt, Maria Becker, Juri Opitz, Vivi Nastase, and Anette Frank. 2019. Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ ISWC*, pages 62–77.

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

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Andrew Pineau. 2013. The abuses of argument: Understanding fallacies on toulmin's layout of argument. *Informal Logic*, 33(4):531–546.

Olesya Razuvayevskaya and Simone Teufel. 2017. Finding enthymemes in real-world texts: A feasibility study. *Argument & computation*, 8(2):113–129.

Stephen Edelston Toulmin. 1958. *The use of argument*. Cambridge University Press.

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

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

David M Williamson. 2013. Probable cause: Developing warrants for automated scoring of essays. In *Handbook of automated essay evaluation*, pages 175–202. Routledge.