Ranking Warrants with Pairwise Preference Learning

Keshav Singh[†] Edwin Simpson[‡] Paul Reisert[*,†] Iryna Gurevych[‡] Kentaro Inui[†,*]

[†]Tohoku University, Sendai, Japan
[‡]Ubiquitous Knowledge Processing Lab (UKP), Department of Computer Science,
Technische Universitat Darmstadt, Darmstadt, Germany
[*]RIKEN Center for Advanced Intelligence Project
{keshav.singh29,inui}@ecei.tohoku.ac.jp
{simpson,gurevych}@ukp.informatik.tu-darmstadt.de
paul.reisert@riken.jp

1 Introduction

The dominant approach to the *ad hoc* evidence detection task has been to establish a pipeline architecture over an initial relevant list of documents followed by different similarity and rule based techniques applied over all candidate pieces of evidence for a given query argument consisting of a claim and topic [1, 11, 9]. Recently, however, it has become questionable as to whether such techniques are adequate in a *'real world'* scenario [3, 10], where evidence needs to be extracted across a diverse range of documents [17]. Such a complex task not only requires evidence detection systems to comprehend the argument but to also analyze why a certain claim follows from its evidence. Instead, current approaches either rely on lexical indicators (e.g., discourse markers such as because, moreover, etc.) or leverage different similarity foundations (LDA threshold, Wordnet sysnet walks, semantic similarity, etc.) which makes current systems unable to distinguish between refuting, supporting, or even invalid evidence.

To overcome this problem, [6] showed that recognizing the implicit link (i.e. warrant) is crucial for understanding why a certain evidence supports any given claim. However, warrants that fill this reasoning gap are implicit in nature, difficult to find explicitly, and vary in reasoning structure [5, 2, 16]. Warrants can also vary widely in quality, particularly if written by crowdworkers [16]. Shown in Figure 1 is an example of claim, evidence and four candidate warrants which try to explicate the notion of supporting evidence relevance. Notice that only W_2 provides sufficient yet necessary explanation and justifies why given evidence supports the claim.

Hence, to further the task of warrant collection, it is intuitive to collect multiple warrants for each claim-evidence pair. We propose the use of crowdsourcing to enable our methodology to scale to larger datasets. To account for the variations in collected warrants, we propose to collect comparisons of warrants from crowdworkers. Our hypothesis is that based on these comparisons, we can devise a methodology to infer which warrant, i.e., reasoning, best bridges the gap between a given claim and evidence. For this we propose to use Gaussian process preference learning (GPPL) [4, 15] to rank multiple warrants for a given claim and evidence pair.

Claim: The use of performance enhancing substances by players is illegal and ethically wrong.
Evidence: President of USA, a former co-owner of Texas ranger, stated that, "Steroids have sullied the game".

Warrants:
- W_1: Performance enhancing substances are used to improve any form of activity performance in humans.
- W_2: Performance enhancing drugs give users an unfair advantage over the rest of the players.
- W_3: The use of anabolic steroids increases the athlete's chance of getting liver cancer.
- W_4: US president is always correct.

Figure 1: An instance of four variable candidate warrants (W_1-W_4) collected via crowdsourcing for a given claim and evidence pair, where W_2 can be considered better reasoning which licenses the move from evidence to claim

To provide annotated data for preference learning, annotators sort lists of warrants according to how well they connect a particular claim with a given piece of evidence. This approach has a number of advantages over classification or annotating numerical quality scores. Firstly, the quality of warrants spans a wide spectrum, so it is unsuitable to assign a small number of class labels, such as "valid" or "invalid", as this does not help us to identify the best warrant for all claim and evidence pairs. Choosing a numerical score directly is known to be harder for annotators than comparing items [7, 8, 18], and annotators may also interpret the values differently to

one another [19].

The sorted lists of warrants are converted to pairwise labels indicating which of each pair of warrants was ranked more highly. The pairwise labels are then provided as input to GPPL, which infers a quality score for each warrant that can be used for ranking. GPPL is able to handle disagreements between workers and is more effective with sparse training data than alternative methods due to its Bayesian approach [14]. Previous work has applied GPPL to ambiguous NLP tasks, including evaluating argument convincingness [14], humour and metaphoricity [13], However, neither GPPL nor any alternative ranking models have previously been applied to warrants.

2 Corpus of ranked warrants

In this section, we describe our proposed method for collecting multiple user-generated warrants for each claim and evidence pair, followed by description of our preference learning method to rank the generated warrants.

2.1 Data

For our experiments, we utilize IBM's Context Dependent Evidence detection (CDED) dataset [11]. The CDED dataset contains 3057 distinct instances covering over 39 different topics. Each instance in CDED consists of (i) topic, (ii) claim, and (iii) a piece of evidence. We randomly sampled 10 of these instances pertaining to different topics and utilized them for our pilot crowdsourcing task.

2.2 Crowdsourcing Tasks

We use Amazon Mechanical Turk (AMT) for crowdsourcing warrants. Although crowdsourcing (CS) is powerful tool for data generation [12], quality control for such a complex task still remains a challenge [6]. Thus we carefully design our task of warrant collection for crowdworkers following [6].

Initially, we provide annotators with set of instructions along with definitions and example annotations. We also emphasize that they write their warrants solely based on their assumptions which are as general as possible and not based on personal experiences.

Warrant generation task Given a claim and evidence piece, workers are required to think and write a basic assumption which they think is necessary to answer why the given evidence supports the given claim. Since this is a complex and challenging task, we give workers the freedom to write any assumption they think of but restrict them from writing false or vague warrants by providing Do's and Dont's guidelines. From this task we collect 10 warrants per claim and evidence pair.

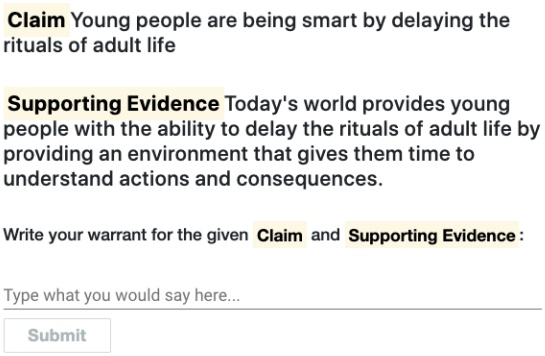

Figure 2: Crowdsourcing interface for collecting warrants. Workers were shown a claim and supporting evidence and asked to create a warrant that links both components.

Preference ranking task After collecting multiple warrants, our next task is to ask crowdworkers to rank these warrants in a simple drag and drop task. The worker is asked to arrange the sequence of the warrants from best(Top) to worst(Bottom) based on their preferences i.e. how well they think the given warrants link the supporting evidence to the claim. For quality control measure, we introduce some dummy warrants and if workers cant rank it appropriately, we automatically reject them.

To discourage noisy annotations, we also warn crowdworkers that their work would be rejected for noisy submissions. We employ simple filtering to exclude crowdworkers who presented copied and pasted claim/evidence as warrants. To see how reasoning varies across workers, we hire 10 crowdworkers per one instance. We hire reliable crowdworkers with 5,000 HITs experiences and an approval rate of 99.0%, and pay $0.20 as a reward per instance.

3 GPPL Model

To apply GPPL, we first convert the ranked list of warrants provided by a crowdworker to a set of pairwise labels. The pairwise labels refer to all pairs of warrants for a claim-evidence pair, and have a value of 1 if the first warrant in the pair was ranked more highly by the crowdworker. Unlike previous applications of GPPL (e.g. [14, 13], we obtain an exhaustive set of pairwise comparisons and do not consider the

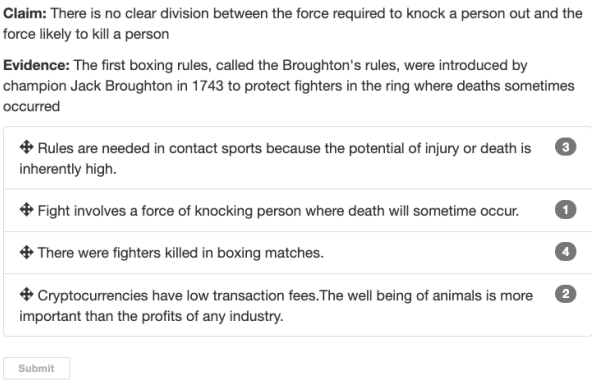

Figure 3: Crowdsourcing interface: Preferences ranking task

features of the warrants, claims or evidences.

Given pairwise labels from multiple crowdworkers for all the claim and evidence combinations, we use GPPL as follows to infer scores for each warrant. Since the value of a warrant depends on the claim and evidence it is being used to connect, our model rates tuples consisting of a warrant, claim and evidence, which we refer to as *instances*.

The GPPL model [4, 15] assumes that each instance, x, has a score, $f(x)$, and that a pairwise label y is chosen by an annotator by comparing instances x_a and x_b with likelihood $p(y = 1|f(x_a), f(x_b)) = \Phi\left(f(x_a) - f(x_b)\right)$, where $y = 1$ indicates that x_a is preferred and Φ is the probit likelihood function. This likelihood allows us to infer the values of $f(x)$ from pairwise labels, assuming that annotators will choose labels less consistently if the values of items x_a and x_b are very similar.

To perform inference, we use stochastic variational inference, a scalable approximate method that can handle large numbers of pairwise annotations and instances [15]. The algorithm returns the posterior distributions over $f(x)$ for all instances in the dataset, which are Gaussian distributions with mean $\hat{f}(x)$. We use these posterior means as estimates of the value of the warrant for linking a particular claim and evidence, and hence to rank the proposed warrants for each claim and evidence pair.

4 Experiments

4.1 Settings

We aggregated the crowdsourced warrants obtained in Section 2 and manually labelled the warrants as good or bad. In total, we labeled 100 warrants obtained from our preliminary crowdsourcing experiment and obtained 54 bad and 46 good warrants.

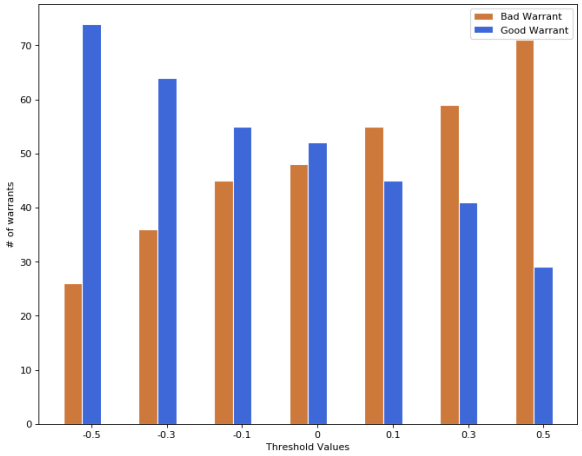

Figure 4: Comparision of GPPL performance on classifying warrants (good, bad) according to different threshold values

We evaluate our GPPL model for binary classification (i.e., whether a warrant is good or bad). We assume that if our model is able to accurately rank the proposed warrants for each claim and evidence pair, then the top (high score) and bottom (low score) ranked warrants should qualify as good and bad warrants, respectively. As an evaluation measure, we report the classification accuracy for different threshold values obtained via warrant scores from our model.

4.2 Results and discussion

As shown in Figure 4, although the GPPL model did not receive supervision about correctly labeled warrants, positive score given by GPPL to warrants matched with the labeled good warrants to some extent and vice versa. As the threshold increases from negative to positive , we found the classification accuracy to increase.

4.3 Qualitative Analysis of Warrants

To evaluate quality of the warrants collected via crowdsourcing, we randomly sampled 40 claim, evidence, and warrant triplets spread across five different topics. We asked two annotators experienced in argumentation mining to score how logical the explanation of the warrants was on a scale of 0-2, where 0 indicates no link and 2 indicates a perfect link between the claim and evidence. We obtained a Krippendorff's α of 0.52, which indicates a moderate to substantial agreement. While the median scoring of both annotators was 1, both scored 2 for 31% and 21% of the instances, respectively, which indicates

that both annotators were able to find considerably good warrants in the randomly sampled data.

5 Conclusion

Towards identifying the implicit link between a claim and evidence pair, we proposed a methodology to infer which warrant best bridges the reasoning gap between claim and evidence by ranking multiple collected warrants via Gaussian process preference learning. Our experiments using simple threshold based classifier have demonstrated that ranking warrants is indeed a challenging and expensive task, but identification of good warrants with the current methodology looks promising.

One immediate future work will be to expand the current crowdsourcing annotation tasks to collect warrants for different topics and automate the ranking method instead of relying on crowdworkers to collect preferences.

References

[1] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[2] 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.

[3] Katarzyna Budzynska and Chris Reed. Advances in argument mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 39–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Wei Chu and Zoubin Ghahramani. Preference learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 137–144. ACM, 2005.

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

[6] 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.

[7] Maurice George Kendall. *Rank Correlation Methods*. Griffin, Oxford, UK, 1948.

[8] David C. Kingsley and Thomas C. Brown. Preference uncertainty, preference refinement and paired comparison experiments. *Land Economics*, 86(3):530–544, August 2010.

[9] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[10] Juri Opitz and Anette Frank. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[12] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866, 2014.

[13] Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Edwin Simpson and Iryna Gurevych. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371, 2018.

[15] Edwin Simpson and Iryna Gurevych. Scalable bayesian preference learning for crowds. *Machine Learning*, 2020.

[16] Keshav Singh, Paul Reisert, Naoya Inoue, Pride Kavumba, and Kentaro Inui. Improving evidence detection by leveraging warrants. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 57–62, Hong Kong, China, November 2019. Association for Computational Linguistics.

[17] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[18] Yi-Hsuan Yang and Homer H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774, May 2011.

[19] Georgios N Yannakakis and John Hallam. Ranking vs. preference: a comparative study of self-reporting. In *International Conference on Affective Computing and Intelligent Interaction*, pages 437–446. Springer, 2011.