Improving Evidence Detection by Leveraging Warrants

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

[†] RIKEN Center for Advanced Intelligence Project [‡] Tohoku University

{keshav.singh29,naoya-i,pkavumba,inui}@ecei.tohoku.ac.jp

paul.reisert@riken.jp

Abstract

Recognizing the implicit link between a claim and a piece of evidence (i.e. warrant) is the key to improving the performance of evidence detection. In this work, we explore the effectiveness of automatically extracted warrants for evidence detection. Given a claim and candidate evidence, our proposed method extracts multiple warrants via similarity search from an existing, structured corpus of arguments. We then attentively aggregate the extracted warrants, considering the consistency between the given argument and the acquired warrants. Although a qualitative analysis on the warrants shows that the extraction method needs to be improved, our results indicate that our method can still improve the performance of evidence detection.

1 Introduction

An argument is composed of two key components: *claim* and *a supporting piece of evidence*. Identification of these components and predicting the relationship among them forms the core of an important research area in NLP known as Argument Mining (Peldszus and Stede, 2013). Although claims can be identified with a promising level of accuracy in typical argumentative discourse (Eger et al., 2017; Stab et al., 2018), identification of a supporting evidence piece for a given claim (i.e., evidence detection) still remains a challenge (Gleize et al., 2019).

Shown in Figure 1 is an example of a given topic and claim, and three evidence candidates from Wikipedia. In this example, identification of the best supporting piece of evidence is challenging, as all three evidence are related to the topic. Although all evidence candidates appear to be semantically similar to the claim, only E_1 supports it, as it has an underlying, implicit link that can be established with the claim (i.e., *children's fundamental right shouldn't be trumped by parental rights*). Thus, for detecting the best piece of evidence for a claim, it is crucial to capture such implicit reasoning between them (Habernal et al., 2018).

Topic: This house believes that male infant circumcision is tantamount to child abuse.
Claim: Infant circumcision infringes upon individual autonomy.
Evidences:
- E_1: In Netherlands, the Royal Dutch Medical Association (KNMG) stated in 2010 that non-therapeutic male circumcision "conflicts with the child's right to autonomy and physical integrity".
- E_2: The British Medical Association states that, "Parents should determine how best to promote their children's interests".
- E_3: American Academy of Pediatrics states that, "Newborns who are circumcised without analgesia experience pain and physiologic stress".

Warrant:
- Children's fundamental right shouldn't be trumped by parental rights.

Figure 1: Three evidence candidates (E_1-E_3) for a given topic and claim, where E_1 can be considered the best evidence piece (shown in blue).

Existing approaches for evidence detection have often relied on lexical features extracted from argument components such as semantic similarity, adjacent sentence relation and discourse indicators (Stab and Gurevych, 2014; Rinott et al., 2015; Nguyen and Litman, 2016; Hua and Wang, 2017). However, no prior work has considered identifying the underlying, implicit reasoning, henceforth *warrants* (Toulmin, 2003), between a claim and a piece of evidence as a means for improving evidence detection. For example, if a model could establish a warrant between the claim and a piece of evidence (e.g., warrant in Figure 1 for E_1), the most plausible evidence piece could be detected.

Towards filling this reasoning gap, Boltužić and

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 57–62
Hong Kong, November 3, 2019.

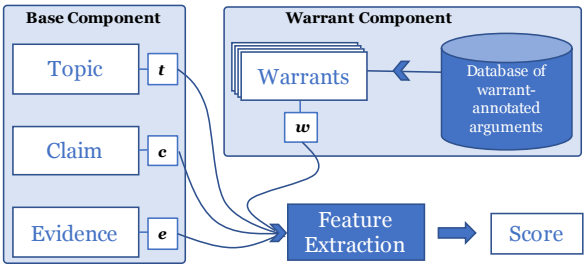

Figure 2: The proposed warrant-aware evidence detection framework.

Šnajder (2016) and Habernal et al. (2018) both created a corpus of explicit warrants for a given claim and its evidence piece. However, to the best of our knowledge, such corpora of explicit warrants have not yet been applied to the task of evidence detection.

In this paper, we explore the effectiveness of leveraging warrants for evidence detection. Given a claim and evidence, our framework first extracts relevant warrants from an existing, well-known corpus of warrant-annotated arguments (Habernal et al., 2018). It then attentively aggregates the acquired warrants, considering the consistency between the given argument and warrants. Our experiments demonstrate that exploiting warrants has the potential to help improve the performance of evidence detection.

2 Proposed Method

2.1 Overview

Given a topic, claim, and a piece of evidence as input, our framework estimates the likelihood of the claim being supported by that evidence piece. As described in Section 1, in order to identify such support relations, it is crucial to recognize the underlying, implicit link between a claim and a given piece of evidence (i.e. warrants). Our framework first extracts multiple warrants that link a given claim to an evidence piece, and later leverages the acquired warrants to estimate the score. We assume that for a given claim and a piece of evidence, there can be several possible variants of warrants for one given claim-evidence pair.

As shown in Figure 2, our proposed framework consists of: (i) *Base Component* and (ii) *Warrant Component*. The Base Component encodes a topic, claim, and an evidence piece into a corresponding vector representation $\boldsymbol{t}$,$\boldsymbol{c}$,$\boldsymbol{e} \in \mathbb{R}^d$. The warrant component then extracts multiple warrants linking the given claim with that piece of evidence and produces its vector representation $\boldsymbol{w} \in \mathbb{R}^d$.

Finally, we generate a feature representation $\boldsymbol{f}$ of all these vectors as follows: $\boldsymbol{f} = [\boldsymbol{t}; \boldsymbol{c}; \boldsymbol{e}; \boldsymbol{w}; \boldsymbol{t} \odot \boldsymbol{c} \odot \boldsymbol{e} \odot \boldsymbol{w}; \boldsymbol{i}] \in \mathbb{R}^{(5d+12d)}$, where $\odot$ denotes element-wise multiplication, and $\boldsymbol{i}$ is the feature vector which captures the pairwise interaction between all ingredients. Analogously to Conneau et al. (2017), we calculate absolute difference and element-wise multiplication for all possible pairs of vectors: $\boldsymbol{i} = \text{concat}(\{[|\boldsymbol{u} - \boldsymbol{v}|; \boldsymbol{u} \odot \boldsymbol{v}] \mid \boldsymbol{u}, \boldsymbol{v} \in \{\boldsymbol{t}, \boldsymbol{c}, \boldsymbol{e}, \boldsymbol{w}\}\})$. Finally, we feed $\boldsymbol{f}$ into a linear classifier: $\boldsymbol{y} = \text{softmax}(U\boldsymbol{f} + \boldsymbol{b})$, where $U \in \mathbb{R}^{(5d+12d)\times 2}$ and $\boldsymbol{b} \in \mathbb{R}^2$ are model parameters to be learned.

2.2 Base Component

The base component produces vector representations of topic, claim, and an evidence piece. This component consists of three types of layers: an embedding layer, a BiLSTM (Hochreiter and Schmidhuber, 1997) layer and a max-pooling layer.

Let $(x_1^t, x_2^t, ..., x_n^t)$ be a sequence of words in a topic. The embedding layer outputs a vector $\boldsymbol{x}_i^t \in \mathbb{R}^g$ for each word x_i^t. The BiLSTM layer then takes a sequence of these vectors $(\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_n^t)$ as an input and produces a contextualized vector $\boldsymbol{z}_i^t = [\overrightarrow{\boldsymbol{h}_i}; \overleftarrow{\boldsymbol{h}_i}]$ for each word, where $\overrightarrow{\boldsymbol{h}_i}, \overleftarrow{\boldsymbol{h}_i} \in \mathbb{R}^h$ are the hidden states of the forward and backward LSTM, respectively. Finally, the max pooling layer extracts the most salient word features over the words to produce a fixed-length vector, i.e. $\boldsymbol{t} = \max_{i=1}^{n}(\boldsymbol{z}_i^t) \in \mathbb{R}^{d=2h}$. In a similar fashion, we obtain vector representations $\boldsymbol{c}, \boldsymbol{e}$ of claim and an evidence piece.

2.3 Warrant Component

Extracting warrants Given a claim and a piece of evidence, our goal is to extract multiple, relevant warrants that link the claim with that evidence piece. As described in Section 1, ideally, we can find plausible warrants for correct claim-evidence pieces but we cannot for wrong pieces. Instead, for wrong claim-evidence pieces, we find non-reasonable warrants that would be less convincing and irrelevant.

Let $\mathcal{D} = \{(t_i, c_i, e_i, w_i)\}_{i=1}^n$ be a database of warrant-annotated arguments, where t_i, c_i, e_i, w_i are a topic, claim, a piece of evidence, and a warrant linking c_i with e_i, respectively. Given an ar-

gument t, c, e to be analyzed, we extract warrants linking c with e via similarity search on $\mathcal{D}$. Specifically, we retrieve the top-m most similar arguments in $\mathcal{D}$ to the given argument in terms of topic, claim and an evidence piece, and then extract warrants from these similar arguments.

We define the similarity between arguments as follows: $\mathrm{sim}(\langle t, c, e\rangle, \langle t_i, c_i, e_i\rangle) = \mathrm{sim}(t, t_i) \cdot \mathrm{sim}(c, c_i) \cdot \mathrm{sim}(e, e_i)$. To calculate the similarity between components u, v, we encode each component into a vector representation $\boldsymbol{u}, \boldsymbol{v}$, and then resort to vector-based similarity. In our experiments, we use Universal Sentence Encoder as a sentence encoder and angular-based similarity as $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v})$, following Cer et al. (2018) because of its state of the art performance in various semantic textual similarity tasks.

Constructing $\mathcal{D}$ is a challenging problem. In our study, we rely on a database of arguments that have arguments which are explicitly annotated with warrants (see Section 3.1 for further details). In future work, we plan to extract warrants from web debate forums, where people frequently discuss controversial topics and ask warrants for discussion with each other.

Encoding warrants Given a set W of extracted warrants $\{w_1, w_2, ..., w_n\}$, we first encode each warrant w_i into a vector representation $\boldsymbol{w_i} \in \mathbb{R}^d$ in a similar manner to topic, claim, and a piece of evidence. Because the quality and relevance of extracted warrants may vary, we attentively aggregate sentence-level vector representations of all extracted warrants. We take a similar approach to Lin et al. (2016), which demonstrated the advantage of sentence level selective attention for multiple sentences, and take advantage of information present in multiple warrants.

Specifically, the final vector representation $\boldsymbol{v}(W) \in \mathbb{R}^d$ is computed as a weighted sum over all warrant vectors:

$$\boldsymbol{v}(W) = \sum_{i=1}^{n} \alpha_i \boldsymbol{w_i}, \tag{1}$$

where α_i is the importance of w_i (s.t.$\sum_{i=1}^{n} \alpha_i = 1$). We calculate α_i as follows:

$$\alpha_i = \frac{e^{f([\boldsymbol{t};\boldsymbol{c};\boldsymbol{e};\boldsymbol{w_i}])}}{\sum_j^n e^{f([\boldsymbol{t};\boldsymbol{c};\boldsymbol{e};\boldsymbol{w_j}])}}, \tag{2}$$

where $f(\boldsymbol{x}) = \tanh(\boldsymbol{u}^\top \boldsymbol{x} + b)$. $\boldsymbol{u} \in \mathbb{R}^{4d}$ and $b \in \mathbb{R}$ are model parameters to be learned. Analogously to attentions in neural models, f estimates the consistency between a given topic, claim, an evidence piece, and warrant.

In our experiments, we also consider a model in which we assume that all warrants are of equal importance and have the same contribution towards the final vector representation $\boldsymbol{v}(W)$, i.e. $\forall i, \alpha_i = \frac{1}{n}$.

3 Experiments

3.1 Dataset

Benchmark of evidence detection To test the model's evidence detection ability, we use the Context Dependent Evidence Detection (CDED) dataset (Rinott et al., 2015). Each instance in CDED consists of (i) topic, (ii) claim, and (iii) a piece of evidence. To create the dataset, Rinott et al. (2015) initially selected 39 topics at random from *Debatabase*.[1] For each topic, they collected 5-7 related Wikipedia articles and then annotated sentences in each article with a claim and its piece of evidence. They also classified each evidence piece into the types *anecdotal*, *study*, and *expert*. In total, the test and training data consists of 3,057 distinct instances (anecdotal: 385, study: 1,020, and expert: 1,896[2]).

Database of warrant-annotated arguments We utilize the dataset of the Argument Reasoning Comprehension Task (ARCT) (Habernal et al., 2018), because it provides a large collection of warrant-annotated arguments that cover a wide variety of topics. The dataset contains 1,970 warrant-annotated arguments covering over 172 topics. Specifically, each instance in the dataset consists of (i) topic, (ii) claim, (iii) premise (i.e., a piece of evidence), (iv) correct warrant, and (v) wrong warrant. For our experiments, we utilize only the correct warrants. The word overlap between topics of CDED and ARCT after stemming and lemmatization was found to be approximately 15%.

3.2 Setting

Evaluation protocol We evaluate our model in the task of *evidence ranking* (Rinott et al., 2015). Specifically, given a claim and candidate evidence, the task is to rank the candidates properly. For each instance in CDED, we extract one false piece of evidence from instances with the same topic but

[1] https://idebate.org/debatabase

[2] Each evidence piece can consists of more than one type.

Number of warrants	Importance α_i	MQ (Anecdotal)	MQ (Study)	MQ (Expert)
None	-	0.47	0.52	0.67
$m = 1$	-	0.48	**0.56**	**0.70**
$m = 1$ (random)	-	0.47	0.51	0.56
$m = 5$	Equal	0.44	0.51	0.65
$m = 5$	Weighted	**0.49**	0.51	0.64

Table 1: Performance of evidence ranking. Results in bold indicate the best MQ score.

different claim. In general, when we have N types of claims in one topic, the task is to rank $N+1$ candidate evidence consisting of one correct and N false evidence. As an evaluation measure, we report Mean Quantile (MQ) score (Guu et al., 2015) which gives a normalized version of Mean Reciprocal Rank. Specifically, for instance, we define the quantile of a correct piece of evidence k as the fraction of incorrect evidence ranked after k. MQ is defined to be the average quantile score over all instances in the dataset, with the quantile ranging from 0 to 1 (1 being optimal).

Following Rinott et al. (2015), we use leave-one-out cross validation schema to evaluate our approach. For every topic, we train our model on instances in all other topics and then test the resulting model on the left out topic. Prior to our experiments, we exclude topics of each evidence type that had less than 3 evidence.

Hyperparameters For both base and warrant components, we use pre-trained 100-dimensional GloVe embeddings (Pennington et al., 2014) to initialize the word embedding layer ($g = 100$). For the BiLSTM layer, we set $h = 100$ (i.e. $d = 200$) and apply dropout before the linear classifier with probability of 0.5. We optimize the categorical cross-entropy loss using Adagrad (Duchi et al., 2011) with a learning rate of 0.01 and the batch size of 32. We choose the model that performs best on the validation set.

3.3 Results

The results are shown in Table 1. The results indicate that incorporating warrant information is effective for ranking evidence across all evidence types. Among warrant-aware models, we found that using a single warrant is more effective overall. We attribute this to the fact that extracted warrants are not of high quality (see Section 3.4), which introduces noisy information into the model. Our future work includes developing a more sophisticated method for extracting warrants. The results also indicate that estimating the importance of each warrant is effective on the anecdotal type evidence.

Type	α	A_1	A_2
Anecdotal	0.50	2.05	2.30
Study	0.50	1.60	2.10
Expert	0.26	1.35	1.95
Overall	0.39	1.60	2.10

Table 2: Results of qualitative evaluation of automatically acquired warrants.

To see the importance of the quality of extracted warrants, we experimented with randomly extracted warrants from the database. The results (i.e. "$m = 1$ (random)") show that the performance does not improve or degrade over the non-warrant-aware model. This indicates that extracting relevant warrants is indeed crucial, and that our improvement is attributed to relevant warrants.

3.4 Qualitative Analysis of Warrants

To investigate the quality of the extracted warrants, two annotators (A_1, A_2) experienced in the field of argumentation were asked to score 20 randomly sampled positive instances for each evidence type. Depending on the degree to which a warrant helped them understand the relation between a claim and a piece of evidence, they were asked to score each instance in the range of 1-5. A score of 1 indicates that the given warrant is unrelated to the evidence piece and its paired claim, and 5 indicates that the relationship between the claim and its piece of evidence pair is easy to understand with the warrant. For calculating the agreement scores, we used Krippendorff's α (Krippendorff, 2011). We also show the average scores given by each annotator.

The results of the analysis are shown in Table 2. Although the anecdotal and study agreement scores can be considered fair, the average scores given by both annotators was low, which indicates that the extracted warrants might not be

as useful in linking the claim to its evidence piece.

One successful example of an automatically extracted warrant is shown in Figure 1. As described in Section 1, a warrant gives good support for the link between the claim and a piece of evidence. Additionally, our framework extracted the warrant "a doctor has a responsibility to treat patients problems at all costs", which does not support the link and is irrelevant.

4 Conclusion and Future Work

In this paper, we have explored an approach for exploiting warrant information for the task of evidence detection. Our experiments demonstrated that leveraging warrants even at the coarse-grained sentence-level can improve the overall performance of evidence detection. However, in our future work, we will focus on a fine-grained level to capture a better reasoning structure of warrants. Furthermore, instead of using separate sentence encoders, we will experiment with using a single general sentence encoder. In our qualitative analysis, we found that the automatically acquired warrants are not of high-quality on average. This can be attributed due to the low lexical overlap between the topics of the two datasets used in our experiments. To address this, we will focus on finding relevant warrants from online web discussion portals, in addition to the current structured database of arguments. Simultaneously, we will explore methods for acquiring warrants at a large-scale, such as crowdsourcing.

References

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing knowledge graphs in vector space. *CoRR*, abs/1506.01094.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208, Vancouver, Canada. Association for Computational Linguistics.

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

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. 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. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. 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. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.