Improving Evidence Detection with Domain-specific Implicit Reasonings

Keshav Singh[†] Naoya Inoue[‡,*] Paul Reisert[♣] Farjana Sultana Mim[†] Shoichi Naito[†,◊,*]
Camélia Guerraoui[†] Wenzhi Wang[†,*] Kentaro Inui[†,*]
[†]Tohoku University [‡]JAIST [*]RIKEN [♣]Beyond Reason [◊]Ricoh Company, Ltd.
{singh.keshav.t4, mim.farjana.sultana.t3, naito.shoichi.t1}@dc.tohoku.ac.jp
{wang.wenzhi.r7, guerraoui.camelia.kenza.q4 }@dc.tohoku.ac.jp
beyond.reason.sp@gmail.com naoya-i@jaist.ac.jp kentaro.inui@tohoku.ac.jp

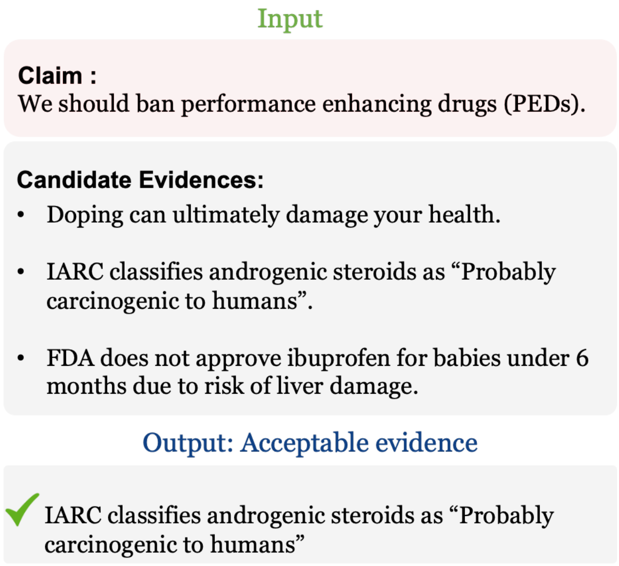
Abstract

Identifying relevant pieces of evidence for a given claim has recently gained significant attention in argument mining due to its downstream use in applications like fact-checking, argument search, etc. While current approaches rely on supervised training of large language models for classifying candidate evidence as acceptable or not for a given claim, they don't generalize well on newer topics for which little to no training data is available. To overcome this issue, in this work, we simultaneously explore the effectiveness of closed-domain approach and leveraging domain-specific implicit reasonings for evidence detection task. Our experimental findings suggest that performance gain in the identification of acceptable evidence for a claim can be further improved even with a small amount of domain-specific implicit reasonings.

1 Introduction

Evidence detection [1] is a sub-task in argument mining that has become an essential component in building natural language systems capable of arguing, debating, and fact checking [13, 9, 2, 10, 17]. Shown in Figure 1, evidence detection refers to the task of identifying evidential statements (i.e., statements of fact, judgement, or testimony) from a set of candidate evidence that support a given claim (i.e., a debatable belief or opinion).

Towards automatically identifying acceptable evidence, recent approaches have relied on pretrained large language models (LLMs) as a default choice because of their outstanding performance in a wide range of NLP tasks [8, 6], including evidence detection [14, 12, 5]. While most of these approaches use supervised learning (i.e., incorporating labeled data for training) and rely on the better generalization ability of LLMs [3, 21], they struggle to produce good results for new topics in which there is little to no training data available. In other words, the quality of their topic generalisation is not adequate [18, 19]. In order to overcome this challenge, in this work, we propose a closed-domain approach towards evidence detection task. [1] Specifically, we follow previous works and take a supervised approach, but instead of directly adopting LLMs for domain-general evidence detection (i.e., train-



Claim :
We should ban performance enhancing drugs (PEDs).

Candidate Evidences:
- Doping can ultimately damage your health.
- IARC classifies androgenic steroids as "Probably carcinogenic to humans".
- FDA does not approve ibuprofen for babies under 6 months due to risk of liver damage.

✓ IARC classifies androgenic steroids as "Probably carcinogenic to humans"

Figure 1: Overview of the evidence detection task we address in this work. Given a claim and a list of candidate evidence, the goal is to identify an acceptable piece of evidence for the given claim.

[1] In this work, the terms *domain* and *topic* share the same meaning, and both refer to the topic of the argument being analyzed.

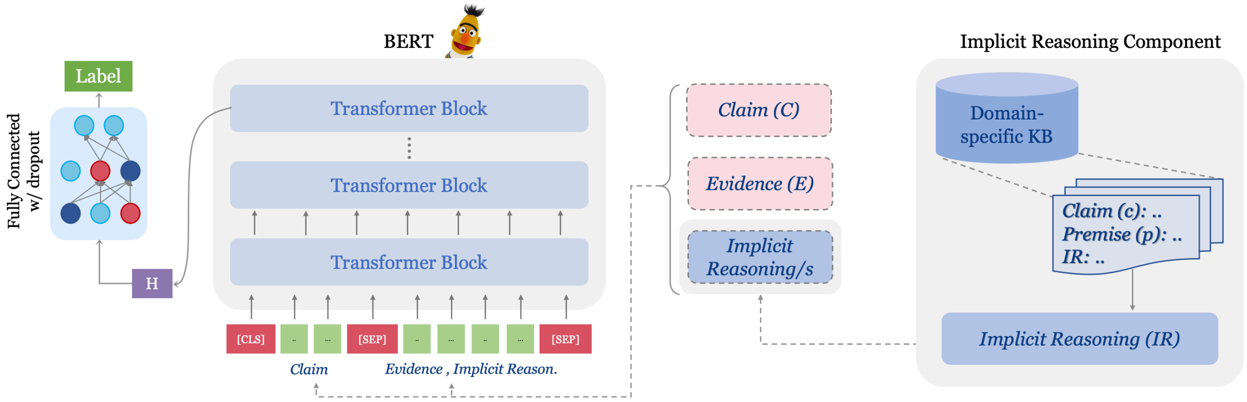


Figure 2: Our proposed framework for the evidence detection task. For a given claim (c) and candidate evidence (e), first the implicit reasoning component extracts relevant implicit reasoning (ir). Later, BERT takes (c), (e) and (ir) as inputs to compute final hidden state of [CLS] token that is fed to a fully-connected feed-forward layer for evidence classification.

ing model on arguments from all topics at once), we train the model on arguments (claim-evidence pairs) belonging to a specific domain along with relevant implicit reasonings (statements that explicitly state the reasoning link between a given claim and evidence) as an input feature. We hypothesize that (i) since LLMs are pre-trained on a large amount of generic text, using a closed-domain approach can assist it to acquire relevant domain-specific knowledge, and (ii) leveraging implicit reasonings belonging to that domain can to be an effective signal for models in establishing the logical link between a given claim and correct evidence candidate [16]. In summary, the contributions of our work are as follows:

- We explore the applicability of a closed-domain approach and domain-specific implicit reasonings towards the evidence detection task and to the best of our knowledge, we are the first to explore this approach.
- We experiment and find that large language models (BERT) trained with domain-specific implicit reasonings in a closed-domain setting performs better than when trained without them.

2 Our Approach

2.1 Overview

Given a topic, claim, and candidate evidence as input, our framework estimates the likelihood of the claim being supported by that candidate evidence. As described in Section 1, we take a closed-domain approach (i.e., we train and test one topic at a time) and simultaneously leverage domain-specific implicit reasonings that are extracted via the implicit reasoning component (See Section 2.2). The complete overview of our evidence detection framework is shown in Figure 2.

Our framework first extracts implicit reasonings (via implicit reasoning component) that link a given claim to an evidence piece, and later leverages the acquired implicit reasoning to estimate the score. We assume that for a given claim and a piece of evidence, there can be several possible variants of implicit reasoning for one given claim-evidence pair.

2.2 Implicit Reasoning Component

Given a claim and a piece of evidence, our goal is to extract relevant implicit reasonings that link the claim with that evidence piece. Ideally, we can find plausible implicit reasonings for correct claim-evidence pieces, but we cannot for wrong pieces. Instead, for wrong claim-evidence pieces, we find non-reasonable implicit reasonings that would be less convincing and irrelevant.

Let $\mathscr{D} = \{(c_i, p_i, r_i)\}_{i=1}^{n}$ be a database of implicit reasoning annotated arguments, where c_i, p_i, r_i are claim, premise and implicit reasoning linking c_i with p_i, respectively [2]. Given a query argument, i.e., claim (c) and candidate evidence (e) to be analyzed, we extract relevant implicit reasonings linking c with e via similarity search on $\mathscr{D}$. Specifically, we retrieve the top-m most

[2] In this work, the utilized source datasets $\mathscr{D}$ of implicit reasonings consists of premise instead of evidence. For more details, refer to [7, 15]

Topic	A	U	Total	A/U
Abolish zoos	22	130	152	0.17
Compulsory voting	12	63	75	0.20
Ban whaling	54	266	320	0.20
Capital punishment	29	199	228	0.15
Legalize cannabis	82	97	179	0.85
School Uniform	10	66	76	0.15
Overall	209	821	1030	0.25

Table 1: Statistics of Evidence data. Here, A and U refer to the number of acceptable and unacceptable evidences for a given topic.

similar arguments in $\mathcal{D}$ to the given query argument in terms of claim and a candidate evidence piece and then extract implicit reasonings from these similar arguments. We define the similarity between arguments as follows: $sim(\langle c, e\rangle, \langle c_i, p_i\rangle) = sim(c, c_i) \cdot sim(e, p_i)$. In our experiments, we use Sentence-BERT [11], a BERT [3] based embedding model shown to outperform other state-of-the-art sentence embeddings methods, to compute the textual embeddings of arguments and calculate semantic similarity between them via cosine-similarity.

3 Experiments

3.1 Source Data

Domain-specific Implicit Reasoning Data As our source of domain-specific implicit reasonings, we utilize the IRAC dataset (Implicit Reasoning in Arguments via Causality) [15], which consists of a wide variety of arguments annotated with multiple implicit reasonings. Overall, the dataset consists of 6 distinct topics covering over 950 arguments that are annotated with 2,600 implicit reasonings. For our experiments, we utilize all 6 topics.

Domain-general Implicit Reasoning Data In order to evaluate the effectiveness of our proposed domain-specific approach, for comparison, we utilize a domain-general corpus of implicit reasonings. Specifically, we rely on the Argument Reasoning Comprehension dataset (ARC) [7], which consists of 1,970 implicit reasoning annotated arguments covering over 172 topics [3]. Each instance in the dataset consists of (i) topic, (ii) claim, (iii) premise, (iv) correct implicit reasoning, and (v) incorrect implicit reasoning. For our experiments, we utilize only the correct implicit reasonings.

Evidence Data Instead of creating a dataset of claim and evidence pairs from nothing, we utilize the IBM-Evidence dataset [4]. Each instance in IBM-Evidence dataset consists of (i) topic (ii) claim and (iii) a piece of candidate evidence, where each candidate evidence is annotated with a score (0-1) indicating its acceptability as evidence for a given claim.

The reason for the selection of this dataset for our experiments is twofold: (i) IBM-Evidence dataset offers 100% coverage of topics present in IRAC dataset. This enables us to adequately test our approach of leveraging domain-specific implicit reasonings for evidence detection task. (ii) IBM-Evidence dataset consists of evidences extracted from Wikipedia articles rather than crowdworkers or experts, hence closely representing real-world evidences. For our experiments, in addition to restricting on 6 topics, we perform an essential pre-processing step and label all candidate evidences as acceptable (score ≥ 0.6) and unacceptable (score ≤ 0.4) in order to classify them. In total, we are left with 1,030 instances of claim-evidence pairs covering 6 distinct topics as shown in Table 1.

3.2 Task Setting

In order to empirically validate the usefulness of utilizing domain-specific implicit reasonings for evidence detection task, we formulate the task in a binary classification setting, where, given a claim (C), a candidate evidence (E) and an implicit reasoning (I), the task is to classify the candidate evidence as acceptable or unacceptable for the given claim.

3.3 Models and Setup

We investigate four different models: (i) a strong baseline model, fine-tuned to classify candidate evidence as acceptable or not, purely based on claim and candidate evidence as input. For this purpose, we select pre-trained BERT model [3], namely **BERT$_{\mathbf{base}}$**, which has been shown to outperform the previously established state-of-the-art on similar tasks [12, 20, 18]. (ii) & (iii) Two separate models to additionally consider the implicit reasonings available via domain-specific or domain-general resource, namely **BERT$_{\mathbf{in}}$**, and **BERT$_{\mathbf{out}}$** respectively. (iv) Addi-

[3] In the original paper, Habernal et al. [7] refers to implicit reasonings as warrants.

Topic	**Random**			**BERT$_{\text{base}}$**			**BERT$_{\text{in}}$**			**BERT$_{\text{out}}$**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Abolish zoos	0.43	0.50	0.47	0.53	0.55	0.53	**0.55**	**0.56**	**0.54**	0.49	0.52	0.50
Compulsory voting	0.42	0.50	0.45	**0.53**	**0.56**	**0.54**	0.50	0.55	0.52	0.51	0.53	0.51
Ban whaling	0.42	0.50	0.45	**0.44**	**0.52**	**0.47**	0.42	0.50	0.46	0.42	0.50	0.45
Capital punishment	0.43	0.50	0.46	**0.58**	**0.57**	**0.55**	0.52	0.54	0.52	0.45	0.51	0.47
Legalize cannabis	0.26	0.50	0.34	0.54	**0.61**	**0.56**	**0.57**	0.56	0.55	0.51	0.57	0.52
School Uniform	0.42	0.50	0.45	0.52	**0.55**	0.52	**0.56**	**0.55**	**0.54**	0.47	0.50	0.48

Table 2: Results of our two baseline models (Random and BERT$_{\text{base}}$) and two implicit reasoning based fine-tuned models (BERT$_{\text{in}}$ and BERT$_{\text{out}}$) in closed-domain setting.

tionally, we consider a random baseline that predicts the most frequent class label as observed in the training data.

3.4 Evaluation Measures

We conduct the fine-tuning experiments for each topic separately and use 70:15:15 splits for training, validation and testing. Since the data for each topic is small (as shown in Table 1), we employ 5-fold cross-validation and average the results. To account for random initialisation of the models, we repeat the experiments with multiple random seeds and report macro-averaged accuracy, precision, recall, and F1 score. In order to address the problem of class imbalance, we calculate class weights to influence the classification of labels during fine-tuning.

4 Results

We evaluate the fine-tuned models for evidence detection on the test set for each topic separately. Note that the results reported consider a single implicit reasoning as input along with claim and candidate evidence. We additionally experimented with multiple implicit reasonings as additional input features but found similar results. As shown in Table 2, all BERT-based models beat the random baseline on all topics, except *Ban whaling*, where their performance is marginally higher. **BERT$_{\text{in}}$** outperforms **BERT$_{\text{out}}$** in all topics except *Ban whaling* and *Compulsory voting*. Overall, **BERT$_{\text{base}}$** outperforms random baseline and achieves higher performance than our implicit reasoning fused models for half of the topics, namely *Compulsory voting*, *Ban whaling* and *Capital punishment*. Our proposed model using domain-specific implicit reasonings i.e., **BERT$_{\text{in}}$** achieved higher performance for only two topics.

4.1 Analysis

Contrary to our expectation, **BERT$_{\text{base}}$** achieved better accuracy than both implicit reasoning fused models on majority of the topics. To better understand this, we analyzed the topic overlap between arguments from ARC and IBM-Evidence dataset and found that arguments on topics *Abolish zoos*, *Ban whaling*, *Capital Punishment* and *School Uniform* were absent in ARC. This explain why **BERT$_{\text{out}}$** performance decreased for these topics. We additionally did manual analysis of implicit reasonings extracted for **BERT$_{\text{in}}$** by randomly sampling 20 instances across all topics and found that only 40% of the extracted domain-specific implicit reasonings were relevant to a given evidence. However, for topics *School Uniform* and *Abolish zoos* they were indeed helpful in finding acceptable evidence.

5 Conclusion and Future Work

In this paper, we explored a closed-domain approach and exploited domain-specific implicit reasonings for the task of evidence detection. Our experiments showed that closed domain approach is beneficial for training large-language models and when leveraging implicit reasonings their performance can improve, given relevant reasonings are available. We hypothesize that reducing the effect of class imbalance with class weights is not sufficient and this might be a possible reason for low performance on topics with severe class imbalance. In our future work, we will focus on utilizing generation models for automatically generating implicit reasonings that can be leveraged for evidence detection task. Simultaneously, we will explore methods for addressing the class imbalance problem.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 22H00524.

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] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In **Proceedings of the first workshop on fact extraction and verification (FEVER)**, pages 85–90, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[4] Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. Corpus wide argument mining—a working solution. In **Proceedings of the AAAI Conference on Artificial Intelligence**, volume 34, pages 7683–7691, 2020.

[5] Mohamed Elaraby and Diane Litman. Self-trained pretrained language models for evidence detection. In **Proceedings of the 8th Workshop on Argument Mining**, pages 142–147, 2021.

[6] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. **arXiv preprint arXiv:2004.10964**, 2020.

[7] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. **arXiv preprint arXiv:1708.01425**, 2017.

[8] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. **arXiv preprint arXiv:1801.06146**, 2018.

[9] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. **ACM Transactions on Internet Technology (TOIT)**, 16(2):1–25, 2016.

[10] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. The evolution of argumentation mining: From models to social media and emerging tools. **Information Processing & Management**, 56(6):102055, 2019.

[11] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 11 2019.

[12] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 567–578, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] 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.

[14] Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pages 599–605, 2018.

[15] Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito, and Kentaro Inui. IRAC: A domain-specific annotated corpus of implicit reasoning in arguments. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pages 4674–4683, Marseille, France, June 2022. European Language Resources Association.

[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] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. An autonomous debating system. **Nature**, 591(7850):379–384, 2021.

[18] Chris Stahlhut. Interactive evidence detection: train state-of-the-art model out-of-domain or simple model interactively? In **Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)**, pages 79–89, 2019.

[19] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. **arXiv preprint arXiv:1904.09223**, 2019.

[20] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The fact extraction and VERification (FEVER) shared task. In **Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)**, pages 1–9, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[21] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. **Advances in neural information processing systems**, 32, 2019.