Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4674–4683
Marseille, 20-25 June 2022

IRAC: A Domain-Specific Annotated Corpus of Implicit Reasoning in Arguments

Keshav Singh[1], **Naoya Inoue**[2,3,*], **Farjana Sultana Mim**[1], **Shoichi Naitoh**[1,3,4] **and Kentaro Inui**[1,3]
[1]Tohoku University, [2]Stony Brook University, [3]RIKEN, [4]Ricoh Company, Ltd.
naoya.inoue.lab@gmail.com, inui@tohoku.ac.jp
{singh.keshav.t4, mim.farjana.sultana.t3, naito.shoichi.t1}@dc.tohoku.ac.jp

Abstract

The task of implicit reasoning generation aims to help machines understand arguments by inferring plausible reasonings (usually implicit) between argumentative texts. While this task is easy for humans, machines still struggle to make such inferences and deduce the underlying reasoning. To solve this problem, we hypothesize that as human reasoning is guided by innate collection of domain-specific knowledge, it might be beneficial to create such a domain-specific corpus for machines. As a starting point, we create the first domain-specific resource of implicit reasonings annotated for a wide range of arguments, which can be leveraged to empower machines with better implicit reasoning generation ability. We carefully design an annotation framework to collect them on a large scale through crowdsourcing and show the feasibility of creating a such a corpus at a reasonable cost and high-quality. Our experiments indicate that models trained with domain-specific implicit reasonings significantly outperform domain-general models in both automatic and human evaluations. To facilitate further research towards implicit reasoning generation in arguments, we present an in depth analysis of our corpus and crowdsourcing methodology, and release our materials (i.e., crowdsourcing guidelines and domain-specific resource of implicit reasonings).

Keywords: argumentation, implicit reasoning, causality, domain-specific resource, logical inference

1. Introduction

Every day, people often engage in different argumentative discourses in written or verbal form (e.g., debates, classroom discussions, or essays). Understanding this kind of discourse requires deducing implicit reasoning (i.e., making logical inferences) between argumentative components, such as the claim and the premise, with information that is not explicitly mentioned (e.g., background knowledge) in the argument (Ennis, 1982; Cain and Oakhill, 1999). For example, consider the argument comprising a claim and its premise, as shown in Fig. 1. Understanding the argument and, henceforth the link between the claim and the premise can be seen as bridging the reasoning gap between them via background knowledge. This process of explicating the reasoning has been shown to help students develop better critical thinking and logical reasoning skills (Erduran et al., 2004). While this process happens relatively quickly and automatically for humans (National Academies of Sciences Engineering and Medicine and others, 2018), a computational system still lacks such a capability due to limited availability of knowledge needed for reasoning and the difficulty in modeling reasoning over such knowledge.

In recent years, significant attention has been given in the field of argumentation mining towards the task of automatic identification and explication of implicit components in arguments (Lawrence and Reed, 2019) because of their importance in downstream tasks such as automatic argument analysis (Hulpus et al., 2019) and educational applications for students in helping them understand and write reasonable arguments (von der Mühlen et al., 2019). Some recent studies have additionally explored the use of a pretrained language models for the explication of implicit reasoning (Becker et al., 2021; Chakrabarty et al., 2021). While this line of research is producing interesting results, the technology has not yet reached the practical level, making it still lacking knowledge and reasoning capability. On the other hand, several previous works have revealed that the innate presence of domain-specific [1] knowledge plays an essential factor in humans that enables them to make reasoning and inferences (Hirschfeld and Gelman, 1994).

Given this background, towards the goal of automatic explication of implicit reasoning, this paper proposes a crowdsourcing-based approach for collecting

Claim: We should *ban surrogacy*.
Premise: Surrogacy often creates *abusive and coercive conditions for women*.

Implicit Reasoning: *Banning surrogacy* causes *decrease in number of women working as surrogates* which suppresses *abusive and coercive conditions for women*.

Figure 1: The implicit reasoning explains the link between the claim and its premise via background knowledge that is useful for understanding the argument.

[*]Present affiliation: Japan Advanced Institute of Science and Technology.

[1]The terms domain-specific and topic-specific are used interchangeably throughout the paper.

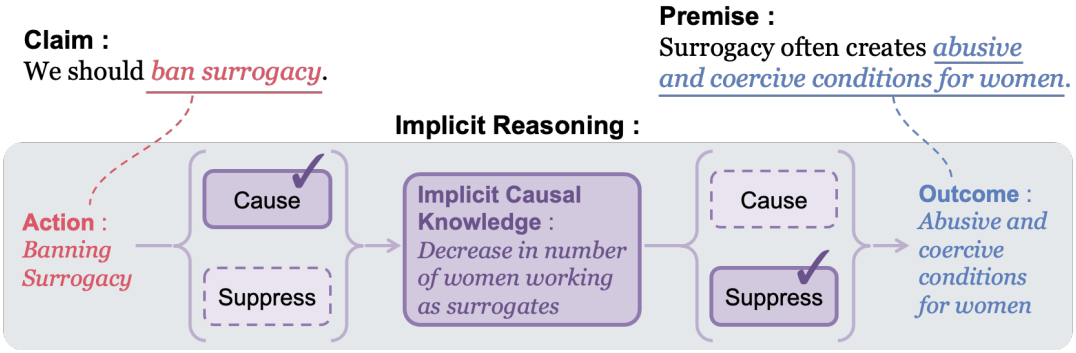

Figure 2: An example of our proposed semi-structured format to explicate implicit reasoning in arguments. Action and outcome represent the key-words/phrases derived from claim and premise respectively. The directed edges between action and outcome are causally linked via implicit causal knowledge, which explains the reasoning link between action and outcome.

domain-specific knowledge to explicate implicit reasoning within a given argument. Specifically, we design an annotation scheme that is applicable for large scale crowdsourcing of implicit reasonings for a given set of claim and premise pairs on a specific topic. The idea is to represent implicit reasoning in a semi-structured format (Fig. 2), where a semi-structured template is used to guide annotators in drawing the inferences between keywords/phrases from a given claim and premise pair. In this annotation scheme, we rely on the notion of causal chains (i.e., cause/suppress labels). It is inspired from the *Argument from Consequences Scheme* (Walton et al., 2008), which has been shown to be useful for explicating implicitly asserted propositions (Feng and Hirst, 2011; Reisert et al., 2018; Al-Khatib et al., 2020; Singh et al., 2021) in arguments. Here, we assume that this protocol can be used for crowdsourcing a collection of domain-specific reasonings for each given argumentative topic and the resulting resource can be incorporated into a model for explicating implicit reasonings for a majority of unseen arguments belonging to that topic. Note that one can consider various potential applications of argument explication where gathering domain-specific knowledge for each topic does make sense. For example, in education, a single topic-specific model can be used by numerous learners and repeatedly year after year, which makes training a model specific to every single topic worthy to consider. For this approach to work, it requires that (i) our approach should be cost-efficient enough for knowledge collection and (ii) the collected knowledge must be effective enough in improving the explication model.

In this paper, we investigate the following questions through a corpus study: (i) Is creating a domain-specific reasoning resource cost-efficient, i.e., can we create a large corpus with reasonable cost and quality? (ii) Can the performance be improved in explicating implicit reasonings when using such a domain-specific resource? Our study positively answers both questions based on a detailed analysis of the quality and cost of collecting implicit reasonings via our methodology.

1. We show that our proposed annotation methodology can be used by non-expert annotators at a reasonable cost while ensuring good quality.

2. We perform empirical evaluation and analysis by leveraging our domain-specific resource for the above task and establish a baseline model for future comparisons.

3. We create and release IRAC(Implicit Reasonings in Arguments via Causality), first domain-specific resource of implicit reasonings for six topics covering 900 arguments annotated with over 2600 implicit reasonings. [2]

2. Related Work

A number of prior works have demonstrated various methods towards the explication of implicit components in arguments ranging from focusing on explicating implicit knowledge to automatically generating implicit reasoning in argumentative texts. Feng and Hirst (2011) were the first to approach this task in computational domain by proposing the use of argumentation schemes (Walton et al., 2008) as a method to capture implicit reasoning in arguments, but no further attempt was made by them in this regard up to this day. Boltužić and Šnajder (2016) hired annotators to fill implicit knowledge in arguments in a domain-general setting, however, they lay no restrictions on their structure and framing, leading them to conclude that the written knowledge pieces heavily vary both in depth and in content.

More recently, Becker et al. (2017) created a corpus of implicit knowledge annotated on top of short German argumentative essays. However, their approach extensively relies on expert annotators, which can be

[2]Our dataset is available at `https://github.com/cl-tohoku/IRAC_2022`

expensive to perform on a large scale. To overcome the prior challenges, Habernal et al. (2018) created a benchmark dataset of domain-general implicit reasonings collected through large scale crowdsourcing with the task of identifying the correct reasoning in a binary classification setting. In contrast to the previous approaches, we focus on a domain-specific approach, where we crowdsource implicit reasonings for multiple arguments for a specific topic and leverage it to train language models to generate implicit reasonings.

At present, the most advanced attempt is from Saha et al. (2021), who created explanation graphs (i.e., ExplaGraphs) to reveal the reasoning process involved in order to explain why a premise supports its claim. They constructed a benchmark dataset that was used to train models to explain the implicit reasoning involved between the argumentative components. While their approach followed a structured representation of implicit reasoning in arguments, the focus of their work was on the model explaining its prediction in a domain-general setting. In contrast to the nature of their study, we propose to collect and utilize domain-specific resource of implicit reasonings that are in semi-structured format, where we focus on causality to explicitly relate the implicit knowledge with key information given in the claim and the premise. Additionally, our corpus contains annotations of implicit reasoning with five times more arguments than the ExplaGraphs, with an average of 150 arguments(each annotated with approximately three implicit reasonings) per topic.

3. Semi-structured Implicit Reasoning

In contrast to explicating implicit knowledge in arguments with general facts or commonsense in unstructured format, we are interested in framing implicit knowledge in the form of argumentation knowledge, which is specifically needed to understand the underlying reasoning link between claim and premise. In particular, as shown in Fig. 2, we develop a template for explicating such implicit reasonings with causality (i.e., *cause/suppress*) and frame its structure in a semi-structured format with the following components:

Action Entity (A): An action entity represents the central objective of the whole argument and is directly derived from the claim as a verbal phrase. This way of framing an action entity from claim is motivated by the conclusion part of the *Argument from Consequences scheme* which states that "Action should/shouldn't be bought about". For example, as shown in Fig. 2, for the claim "We should ban surrogacy", the action can be framed as *"Banning surrogacy."*

Outcome Entity (O): An outcome entity represents the consequence of doing an action, where the consequence is either caused or suppressed by the action. The outcome entity is directly derived from the premise with slight modifications in its phrasing. For example, as shown in Fig. 2, for the premise "Surrogacy often creates abusive and coercive conditions for women", the outcome can be framed as *"Abusive and coercive conditions for women,"* such that it forms the following relation: *"Banning surrogacy"* $\xrightarrow{\text{suppress}}$ *"Abusive and coercive conditions for women."*

Implicit Causal Knowledge (I): In order to understand why/how the premise offers support to the claim, we need to explicate knowledge that is either missing or implicit in the argument. Specifically, we need knowledge that explains the causal connection between the action and outcome entities such that the reasoning link between the claim and the premise becomes clear. For example, the implicit knowledge, i.e., *"decrease in number of women working as surrogates"* (as shown in Fig. 2), is required to understand why/how banning surrogacy suppresses abusive and coercive conditions for women. We term such knowledge as *implicit causal knowledge* and represent it along with the action and outcome entities in the following form:

- Banning surrogacy $\xrightarrow{\text{cause}}$ *Decrease in number of women working as surrogates.*
- *Decrease in number of women working as surrogates* $\xrightarrow{\text{suppress}}$ Abusive and coercive conditions for women.

Causal Relation: The causality between the action entity, the outcome entity and the implicit causal knowledge is represented with *cause/suppress* labels. Although, the expressible quality of the implicit reasoning will be reduced by employing predefined causal labels, we hypothesize that majority of typical instances of implicit reasoning in arguments can be captured by encoding such causal labels.

Figure 2 shows the final implicit reasoning representation in a semi-structured format along with the other aforementioned components.

4. Crowdsourcing Semi-structured Implicit Reasoning

We design a two-phase annotation process to obtain high-quality semi-structured implicit reasonings on a large scale, where each phase (§ 4.1 and § 4.2) can be operated through crowdsourcing on Amazon Mechanical Turk (AMT). In Phase 1, we describe how to obtain the main components that are required to frame the implicit reasoning (§ 3). In Phase 2, we verify the correctness of the collected implicit reasonings and refine them if necessary.

Source Data Instead of collecting the initial claim and premise pairs from scratch, we utilize a well-known dataset of debatable arguments, IBM-30K corpus (Gretz et al., 2019), for our annotation task. The reason for our choice of IBM-30K is as follows.

First, it already consists of arguments in the form of claim and premise for multiple debatable topics that were collected actively from annotators with strict quality control measures as opposed to being extracted

from targeted audiences such as debate portals. This represents a vast majority of all the possible arguments that can be made for a given topic.

Second, we assume that annotation of implicit reasoning on top of the arguments collected by annotators might be highly feasible as it more or less reflects how majority of people make arguments, i.e., often a lot of information in arguments is left implicit.

Third, since the dataset is already available and can be extended to include additional topics, we believe that this will help us to extend our domain-specific resource of implicit reasonings easily.

We select a subset of six common debatable topics out of a total of 71 topics in IBM-30k for our implicit reasoning annotation task. We filter arguments of low point-wise quality (below 0.5) and unclear stance (below 0.6) to make sure that arguments of sufficient quality are used for our annotation task. After the filtering steps, 952 arguments were yielded for the six topics, which we use for our crowdsourcing tasks.

4.1. Phase 1: Framing Implicit Reasoning

In order to frame semi-structured implicit reasoning, we need four main components (§ 3), i.e., *action entity*, *outcome entity*, *implicit causal knowledge*, and *causal relations*. Specifically, for a given claim, premise and action entity, the annotator is asked to derive the outcome entity (STEP 1) and frame the implicit reasoning by annotating other components (STEP 2). In this phase, we allow a maximum of five annotators to write implicit reasoning per claim and premise pair.

Deriving Action Entity (A) We obtain action entity from its corresponding claim by automatically deriving it as a verbal phrase through a simple rule-based matching via spaCy (Honnibal et al., 2020). For example, the action entity *"Introducing compulsory voting"* can be derived from the claim *"We should introduce compulsory voting."*

Deriving Outcome Entity (O) We leverage crowdsourcing to derive the outcome entity from the premise. We assume that there can be multiple ways one can phrase an outcome entity as a consequence of doing an action and such diversity can result in different implicit reasonings. For example, for the following claim and premise:

(1) **Claim**: We should abolish intellectual property.
Premise: People or companies owning the rights to certain ideas can create a closed market, where the owners of such ideas are able to set the price without the fear of competition.

There can be more than one way to derive outcome entity and annotate the relation between action and outcome entity: (i) *Abolishing intellectual property rights* $\xrightarrow{\text{suppress}}$ *Creation of a closed market* and (ii) *Abolishing intellectual property rights* $\xrightarrow{\text{cause}}$ *Fear of competition*, which may consequently result in different implicit reasonings. An example annotation via our crowdsourcing interface is shown in Fig. 3, where in Step 1 annotators are asked to derive the outcome entity for a given premise [3].

Annotating Implicit Causal knowledge (I) In this step, we assume that annotation of such knowledge may not be possible for every claim and premise pair. Specifically, for a bad premise, there may be no feasible way to explicate any causal knowledge that links a claim to its premise. For example, given a claim: *"We should introduce a multiparty system"* and a premise: *"Introducing a multiparty system is the right thing to do,"* it is not possible to write any implicit causal knowledge since the argument is a fallacy (i.e., begging the question), where premise provides no adequate support to the claim. Similarly, for arguments with very good premise, it may not be necessary to annotate any implicit causal knowledge since it might already be explicated in the premise. In order to handle such cases, prior to Step 2, we explicitly ask annotators to judge the feasibility of annotating implicit causal knowledge for a given action entity and their derived outcome entity (see "Question" in Fig. 3). This is a challenging step as annotators may be biased to answer "No" or "Unsure" to avoid doing the task and complete the task quickly. To avoid this issue and reduce biased annotations, we treat this as a bonus question and grant bonus depending on the majority responses, i.e., if majority of the annotators annotate implicit causal knowledge for a given claim and premise, a bonus is granted to the majority and vice versa.

An example annotation for Step 2 is shown in Fig. 3, where annotators are provided with a predefined template for constructing the relationship between action entity, outcome entity, and implicit causal knowledge along with causal relations. Instead of framing the template as a single chain, we rephrase it into individual relations as: (i) *Action Entity* $\xrightarrow{\text{cause/suppress}}$ *Implicit Causal Knowledge* and (ii) *Implicit Causal Knowledge* $\xrightarrow{\text{cause/suppress}}$ *Outcome Entity*.

Annotating causal relations As shown in Fig. 3, the annotation of causal relations between components is done alongside the annotation of implicit causal knowledge. Annotators are asked to pick one out of two choices of causal relations (i.e., cause and suppress) to form the causal connection between *(action entity and implicit causal knowledge)* and *(implicit causal knowledge and outcome entity)*. We include additional sanity checks with the final annotated implicit reasoning for annotators to confirm their annotation.

[3] We avoid using complicated jargon in our crowdsourcing interface in order to make the task easier for annotators to understand. We found this to produce better annotations and fewer errors by non-expert annotators. Specifically, we refer to the claim as stance, premise as supporting statement, implicit causal knowledge as intermediate knowledge, causal relations as connectors and implicit reasoning as logical flow.

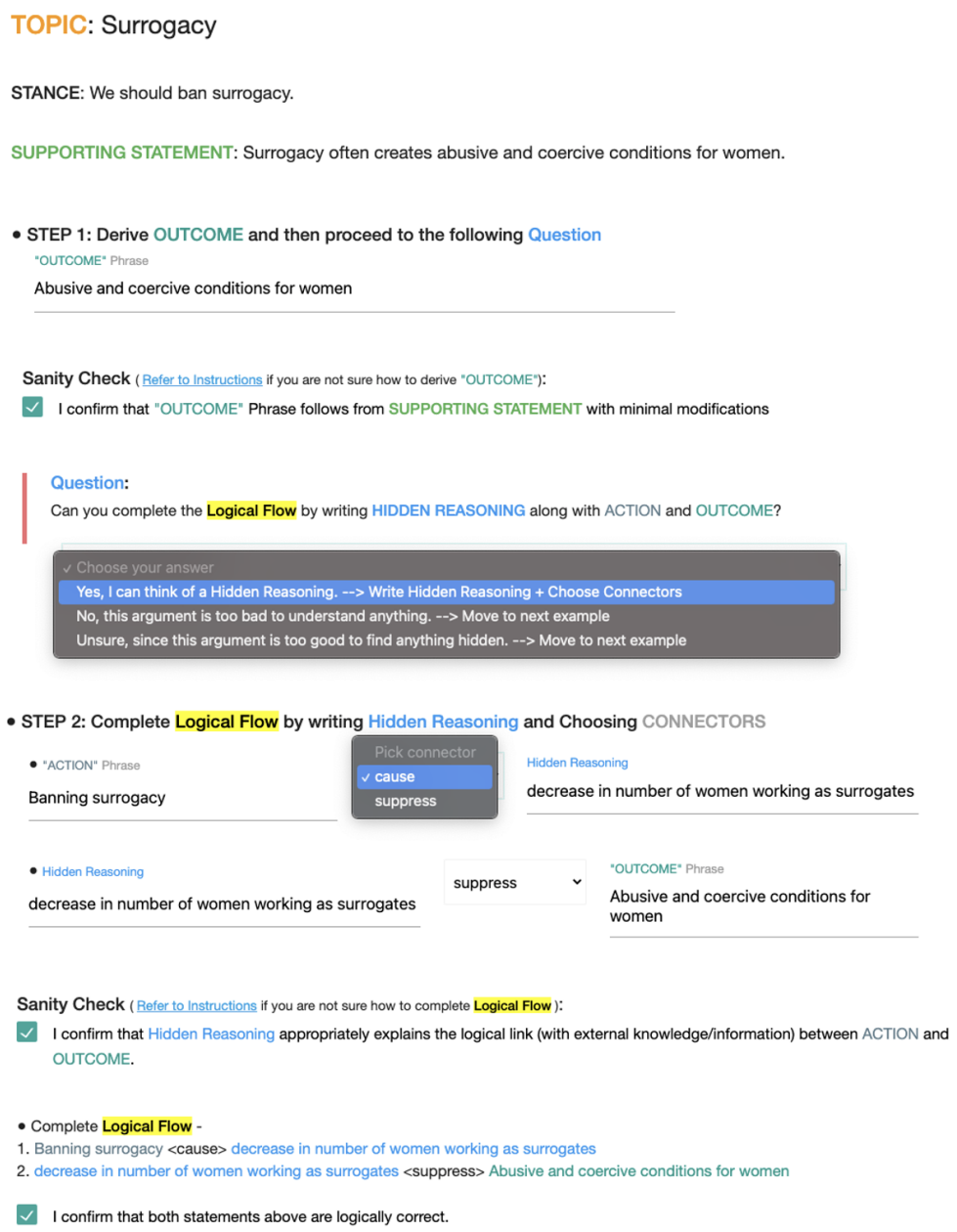

Figure 3: The interface of our crowdsourcing task for Phase 1. This phase consists of two steps, where STEP 1 is mandatory while STEP 2 depends on the choice made by crowdworkers for the Question preceding STEP 2.

4.2. Phase 2: Correctness Verification

Prior to designing this phase, we manually analyzed a fraction of all the implicit reasonings collected in Phase 1. We also asked experts, who are researchers in argumentation, to judge the correctness of the annotations and asked their opinion on the criteria on which implicit reasonings can be evaluated. Overall, the manual analysis showed that 70% of annotations were correct, and based on expert comments and observations, we design Phase 2 to further filter the collected annotations.

Given the implicit reasoning collected in Phase 1, we leverage crowdsourcing to verify their correctness in three distinct criteria: (i) logical correctness, (ii) implicit causal knowledge correctness, and (iii) keyword correctness.

We allow a maximum of three annotators to judge the correctness of an implicit reasoning where each one is asked to verify if the implicit reasoning fulfills each criterion or not. For each annotator, an implicit reasoning is considered correct if and only if it passes all the three criteria; otherwise, it is considered incorrect. We took majority voting, which means if 2/3 of the annotators thought it was incorrect, we mark it as incorrect and do not include it in our final dataset. To make the implicit reasoning coherent and readable for the annotators,

we frame the implicit reasonings as a concatenated structure of all the previous components as follows: *(A) cause/suppress (I). And (I) cause/suppress (O).*

Logical Correctness Following the previous study on the logical quality of arguments (Johnson and Blair, 2006; Wachsmuth et al., 2017), here, we verify the deductive validity of our annotated implicit reasonings. Specifically, given an implicit reasoning, we ask annotators to infer through it such that the implicit causal knowledge component logically follows from the preceding action entity and enables deduction of the given outcome entity.

Implicit Causal Knowledge Correctness For the implicit reasoning to be correct, it is necessary for the implicit causal knowledge to act as intermediate link between keywords from the claim and the premise. In case it is paraphrased from the premise, incoherent, or introduces irrelevant knowledge between action and outcome entity, the implicit causal knowledge is considered incorrect.

Keyword Correctness The derived keywords from the premise (i.e., outcome entity) play an important role in framing the implicit reasoning. As such, to fulfill this criteria, the keywords must be coherent and convey the same semantic meaning as stated in the premise; otherwise, the annotated implicit reasoning cannot be treated correct due to the semantic differences between actual premise and derived outcome entity.

4.3. Pilot Phase

Prior to conducting the main crowdsourcing of implicit reasonings, we conduct multiple annotation studies and pilot runs on AMT to finalize our crowdsourcing design. Since our annotation task is comparatively challenging and non-expert annotators might find it difficult, we successively discussed and refined the task design and instructions by consulting with experts, and taking into account their comments and suggestions. In order to address any ethical issues (Adda et al., 2011) raised by our task, we actively monitor the feedback given by the annotators and communicate with them to resolve any questions/comments raised. In order to further adapt the task to non-expert annotators, we manually verified their annotations after each change in pilot run and provided them with constructive feedback to assist them in understanding the tasks as well as improve the quality of annotation. We found this strategy to work the best in terms of end quality annotations as well as simplifying the task. All the annotators who performed our task were paid in accordance with the minimum wage which was calculated based on their average work-time (See Appendix for further details).

5. IRAC dataset

5.1. Statistics

In Phase 1, we collect a total of 3569 implicit reasonings for 952 claim and premise pairs covering six debatable topics. While in Phase 2, we verify all the collected implicit reasonings and are left out with 2636 implicit reasonings for 909 claim and premise pairs. An average of about three implicit reasonings per claim and premise pair were found to be annotated. Out of 2636 annotations, a total of 2617 implicit reasonings and 2,200 implicit causal knowledge were found to be unique. This shows that similar implicit causal knowledge can be applied to different claim and premise pairs. Table 1 shows additional statistics on (i) the number of implicit reasoning annotations for claim and premise pairs per topic; (ii) the coverage, i.e., % of claim and premise pairs with annotated implicit reasonings per topic; and (iii) the average number of implicit reasonings per claim and premise pair. As shown in Table 1, 95% of the claim and premise pairs in IRAC dataset contain at least one annotated implicit reasoning and 83% of them have at least two annotated implicit reasonings. This indicates that most of the claim and premise pairs can be annotated with implicit reasoning, i.e., our annotation methodology results in high coverage of implicit reasonings for a given set of claim and premise pair. This observation further supports our initial assumption of feasibility of annotating implicit reasonings on top of the IBM-30K arguments with causality.

We create our final argumentative dataset of 2,636 implicit reasonings that are annotated for 909 claim and premise pairs via causality (IRAC) covering six topics. Example annotation from our final curated dataset is shown in Table 2, where the implicit reasoning between claim and premise is made explicit by inserting the **implicit causal knowledge**: *"all people to be mandatorily required to voice their opinions by voting"* and causal labels between **action entity** and **outcome entity**. In total, we discarded 43 claim and premise pairs at the end of Phase 2 as no implicit reasoning could be annotated for them or the annotated implicit reasonings were not correct. We manually analyzed such instances and found that these claims had premises which were either too good or bad to come up with any implicit reasoning.

5.2. Quality analysis

As our dataset only consists of implicit reasoning that were labeled as correct by annotators via majority voting, we apply additional steps to verify the crowdsourced annotations. We ask two experts to repeat the same process as explained in Phase 2. The experts were given 50 implicit reasoning randomly sampled from IRAC dataset and were asked to label the implicit reasoning for a given claim and premise as either correct or incorrect. We measure the agreement between the two experts via Krippendorff's α (Krippendorff, 2011). After aggregating experts annotation, we obtain an Krippendorff's α of 0.64, where the first expert labeled 38 while the second expert labeled 34 implicit reasonings as correct. This shows that our non-expert

Topic	# Claim-Premise	# Implicit Reasonings	IRs ≥ 1	IRs ≥ 2	Avg. # Implicit Reasonings per Premise
Abandon use of school uniform	145	483	99%	95%	3.3 (144)
Abolish capital punishment	176	322	86%	60%	2.1 (152)
Abolish zoos	141	390	98%	86%	2.8 (139)
Ban whaling	164	468	96%	83%	3.0 (158)
Introduce compulsory voting	116	376	100%	94%	3.2 (116)
Legalize cannabis	210	597	95%	86%	2.9 (200)
Total	952	2636	95%	83%	2.9 (909)

Table 1: Statistics of IRAC dataset. IRs ≥ 1 and IRs ≥ 2 denote the percentage of claim and premise pairs with at least one and at least two annotated implicit reasonings, respectively.

Claim	We should introduce compulsory voting.
Premise	Everybody has the responsibility to give their opinion on what happens in their country.
Implicit Reasoning	*Introducing compulsory voting causes* ***all people to be mandatorily required to voice their opinions by voting*** *causes everybody giving their opinion on the issues in their country.*

Table 2: Example annotation of implicit reasoning that links the claim and premise, comprising implicit causal knowledge (in bold) linked with action and outcome entities.

annotators did a fairly good job on the task of annotating as well as verifying the correctness of final implicit reasonings.

6. Experiments

6.1. Task setting

In order to empirically validate the usefulness of our domain-specific resource (IRAC) for explicating implicit reasoning, we utilize it to tackle the following domain-specific generative task: given a claim and its premise (*C*, *P*) on a specific topic, generate the implicit reasoning (*R*). The generated implicit reasoning must explicate the intermediate implicit causal knowledge, such that it links the keywords from *C* and *P* with appropriate causal labels.

6.2. Setup

For establishing a strong baseline, we assume that if such a domain-specific resource is not available, then pre-trained language models (LM) might be the best option to generate implicit reasonings. However, any vanilla pre-trained LM might not be familiar with this task, so we adapt them to this specific task setting so as to teach the format of the task to any kind of models. Hence, we propose to use out-of-domain instances to adapt a given LM to this task (i.e., using instances belonging to a variety of different topics), which we then use as our strong baseline. Consequently, we compare the usefulness of our in-domain (i.e., domain-specific) resource on top of this strong baseline.

In summary, we evaluate the task in two separate settings: (i) **Out-of-domain setting**: As our baseline, we utilize a pre-trained language model (LM) and finetune it in an out-of-domain setting. Specifically, we finetune the LM on all instances from all topics expect one and test the fine-tuned model on the left out topic. (ii) **In-domain setting**: For empirically verifying the performance gain with domain-specific resource, we finetune the LM on training instances from one topic and test the fine-tuned model on the same topic with 80:20 train-test split. We report the final results as average score of fivefold cross-validation runs.

Evaluation Measures We use the BLEU metric (Papineni et al., 2002), one of the most widely used automatic metrics for generation tasks to compute BLEU-1 (B1) and BLEU-2 (B2) scores between our model's output and the human annotated implicit reasonings. We also report F1-Score (BS) of BERTScore (Zhang et al., 2019), which is a metric for evaluating text generation using contextualized embeddings. We evaluate the results by only considering the generated implicit causal knowledge as inclusion of action entity and outcome entity may lead the automatic metrics to give a higher score. This is due to the fact that action entity is similar throughout the topic and outcome entity can be very similar if not same. Hence, during each setup, we trim the generated implicit reasoning to contain only implicit causal knowledge.

6.3. Models

Following the previous works on implicit knowledge generation, we carried out an experiment with BART (Lewis et al., 2019), which is a type of generative LM, in each of our task setting. BART (Lewis et al., 2019) is a pre-trained conditional language model that combines bidirectional and autoregressive transformers. It is implemented as a sequence-to-sequence model with a bi-directional encoder over corrupted text and a left-to-right autoregressive decoder. We use the pre-trained version of BART model provided by HuggingFace Transformers library (Wolf et al., 2020) and

Topic	Baseline			Our Model		
	B1	B2	BS	B1	B2	BS
	Out-of-Domain			In-Domain		
Zoos	0.21	0.04	0.16	0.44	0.28	0.37
Whaling	0.16	0.03	0.20	0.40	0.24	0.37
Cannabis	0.33	0.10	0.19	0.45	0.21	0.48
Voting	0.19	0.07	0.23	0.38	0.21	0.36
School uniform	0.23	0.04	0.27	0.36	0.17	0.41
Capital punishment	0.16	0.02	0.18	0.17	0.03	0.16

Table 3: Automatic evaluation of implicit reasoning (generation by fine-tuned BART) in two settings based on BLEU1 (B1), BLEU2 (B2) and BERTScore (BS).

Claim	We should legalize cannabis.
Premise	Legalizing cannabis can help people with certain health problems be relieved of their symptoms.
Implicit Reasoning	
Gold	*Legalizing cannabis causes **easy access to the drug for the needy** causes helping people with certain health problems be relieved of their symptoms.*
In-domain	*Legalizing cannabis causes **extensive medicinal research on cannabis** causes relief in health problems.*
Out-of-domain	*Legalizing cannabis causes **good medicinal use** causes relieve of patients symptoms.*

Table 4: Example of implicit reasonings generated for a given premise and claim by BART fine-tuned in in-domain and out-of-domain settings. Text in bold depicts how our fine-tuned models explicate and adapt implicit causal knowledge to make inference between claim and premise.

fine-tune it on our corpus.

Fine-tuning To fine-tune BART, we give concatenated C and P as input sequences to the encoder, whereas encoded R is given as labels to the decoder part of BART. Accordingly, our labeled sequences given to decoder part of BART are structured as follows: "*A <causal label> I. And I <causal label> O*", where A is the action entity, I is the implicit causal knowledge, O is the outcome entity, and <*causal label*> can be either one of *cause* or *suppress*. During inference, for a given input sequence, we only focus on reconstructing the complete sequence as given to the decoder. We also experimented with using special delimiter <SEP> to assist model to better differentiate between C, P, and I, but this did not yield good results possibly due to smaller number of training instances.

6.4. Results

As shown in Table 3, of all topics, BART fine-tuned on IRAC in the in-domain setting yields the best results while performs worse in the out-of-domain setting. We also note that fine-tuned BART in both settings generates syntactically correct implicit reasonings; however, out-of-domain fine-tuning generates implicit reasonings that are either incorrect or nonsensical. Examples of generated implicit reasonings via each setting are shown in Table 4.

We manually analyze 100 randomly selected implicit reasonings, each generated by fine-tuned BART in out-of-domain and in-domain settings. Similar to Phase 2, we hired annotators from AMT platform and asked them to judge the correctness of the generated implicit reasoning, i.e., binary classification where annotators had to mark it as correct or incorrect. Each implicit reasoning was judged by three annotators. After considering majority voting, for out-of-domain setting based generation, 56% of instances were marked correct, while for in-domain-based generation, 72% of instances were verified to be correct. Additionally, we manually analyzed the implicit reasonings generated via each setting and notice that for both the settings, the model generated mostly repetitive implicit causal knowledge for numerous instances for the topic: *"We should abolish capital punishment,"* which might be due to less number of training instances available for the topic. To further investigate it, we repeat the experiments with different input prompt, for example, "*A <causal label> I which <causal label> O*" but find no improvement in the results.

7. Conclusion and future work

We propose and create a domain-specific approach to explicate implicit reasonings in arguments and create a dataset of 2,636 implicit reasonings for six topics. We carefully design the annotation framework and show that non-expert annotators can perform the quality annotations and such a dataset can be created at a reasonable cost. Finally, we leverage our corpus to automatically generate implicit reasonings and empirically evaluate the performance gain of language model fine-tuned on our dataset. Our model that is fine-tuned on IRAC in the in-domain setting outperforms the baseline model trained in the out-of-domain setting, which further shows the importance of domain-specific resource, and we believe future research in this direction is a worthwhile effort. In the future, we would like to expand the current corpus to include additional topics as well as the size of the current corpus to include more arguments and annotated implicit reasonings. Additionally, we would like to investigate the effect of using domain-specific resource on top of currently available domain-general resources in the task of implicit reasoning generation.

8. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 22H00524 and NEDO JP1234567. We would like to thank the members of Tohoku NLP lab for their insightful feedback, and the experts and annotators for their valuable time and effort.

9. Bibliographical References

Adda, G., Sagot, B., Fort, K., and Mariani, J. (2011). Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use. In *5th Language and Technology Conference*.

Al-Khatib, K., Hou, Y., Wachsmuth, H., Jochim, C., Bonin, F., and Stein, B. (2020). End-to-end argumentation knowledge graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7367–7374.

Becker, M., Staniek, M., Nastase, V., and Frank, A. (2017). Enriching argumentative texts with implicit knowledge. In *International Conference on Applications of Natural Language to Information Systems*, pages 84–96. Springer.

Becker, M., Liang, S., and Frank, A. (2021). Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online, June. Association for Computational Linguistics.

Boltužić, F. and Šnajder, J. (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, August. Association for Computational Linguistics.

Cain, K. and Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and writing*, 11(5):489–503.

Chakrabarty, T., Trivedi, A., and Muresan, S. (2021). Implicit premise generation with discourse-aware commonsense knowledge models. *arXiv preprint arXiv:2109.05358*.

Ennis, R. H. (1982). Identifying implicit assumptions. *Synthese*, pages 61–86.

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

Feng, V. W. and Hirst, G. (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, June. Association for Computational Linguistics.

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2019). A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*.

Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hirschfeld, L. A. and Gelman, S. A. (1994). *Mapping the mind*. Citeseer.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hulpus, I., Kobbe, J., Meilicke, C., Stuckenschmidt, H., Becker, M., Opitz, J., Nastase, V., and Frank, A. (2019). Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ ISWC*, pages 62–77.

Johnson, R. H. and Blair, J. A. (2006). *Logical self-defense*. Idea.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, December.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

National Academies of Sciences Engineering and Medicine and others. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Reisert, P., Inoue, N., Kuribayashi, T., and Inui, K. (2018). Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*. Association for Computational Linguistics, November.

Saha, S., Yadav, P., Bauer, L., and Bansal, M. (2021). Explagraphs: An explanation graph generation task for structured commonsense reasoning. *arXiv preprint arXiv:2104.07644*.

Singh, K., Mim, F. S., Inoue, N., Naito, S., and Inui, K. (2021). Exploring methodologies for collecting high-quality implicit reasoning in arguments. In *Proceedings of the 8th Workshop on Argument Mining*, pages 57–66, Punta Cana, Dominican Republic,

November. Association for Computational Linguistics.

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

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

Crowdsourcing details Based on our findings from the pilot tests, we only allow annotators who have $\geq$ 98% acceptance rate and $\geq$ 5,000 approved human intelligence tasks for our main annotation tasks (i.e., Phase 1 and Phase 2). Prior to each main task, we additionally hold a preliminary qualification quiz that consists of ten basic questions for testing the annotators' ability to differentiate between implicit and explicit knowledge in a given argument. Workers who score more than a pre-defined threshold ($\geq$ 80%) are granted access to do our tasks. In total, 51 workers who cleared the qualification quiz were selected for Phase 1, and 76 workers were selected for Phase 2. We took additional measures to make sure that annotators from Phase 1 and Phase 2 did not overlap.

Cost Breakdown The annotators were paid according to the minimum wage \$12/hr (\$0.45 for Phase 1 and \$0.20 for Phase 2) during the pilot as well as during main crowdsourcing, which is calculated by conducting many trials and based on their average worktime to ensure fair pay. A separate set of 47 workers in total were selected for bonus pay due to their high quality work. The cost of conducting pilot tests were about \$210 for Phase 1 and \$250 for Phase 2. Separate bonus of \$600 was given to workers who did the task exceptionally well and provided valuable feedback. In total, the cost of creating the final corpus was approximately \$3500 excluding cost of pilot runs. For each topic, the overall cost of annotating implicit reasonings was in the range of \$550 to \$700 for about 150 arguments on average. The total costs for our crowdsourcing tasks were about \$4690 including bonuses, pilot-runs and fees paid to the AMT platform.