博 士 学 位 論 文

論文題目　Bridging Implicit Reasoning Gap in Arguments: An Overnight Approach

提　出　者　　東北大学大学院情報科学研究科

システム情報科学　　専　攻

学籍番号　　C0ID2006

氏　名　　シング　ケシャワ

Bridging the Implicit Reasoning Gap in Arguments: An Overnight Approach

Keshav Singh

Department of System Information Sciences
Graduate School of Information Sciences
Tohoku University
Sendai, Japan

This dissertation is submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

March 2023

Supervisor: **Professor Kentaro Inui**
Natural Language Processing Laboratory,
Department of System Information Sciences,
Graduate School of Information Sciences,
Tohoku University

Examiners: **Professor Shinichiro Omachi**
Department of Communications Engieering,
Graduate School of Engineering,
Tohoku University

Professor Xiao Zhou
Graduate School of Information Sciences,
Tohoku University

Professor Jun Suzuki
Center for Data-driven Science and Artificial Intelligence (CDS),
Tohoku University

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Professor Kentaro Inui, for his invaluable guidance, support and encouragement throughout my PhD journey. His profound knowledge, insightful suggestions and constructive criticism have been instrumental in shaping my research and helping me grow as a researcher. I am deeply grateful for the time he has dedicated to advising me and for his unwavering faith in my abilities.

I would also like to extend my heartfelt thanks to Professor Jun Suzuki, who has always been a source of inspiration and support. His invaluable insights and expertise have greatly enriched my research and contributed to the success of this dissertation. I am grateful for his guidance and encouragement, which have helped me stay focused and motivated throughout my PhD.

I am deeply indebted to my mentors, Dr. Naoya Inoue (JAIST), Dr. Paul Reisert and my colleague Dr. Farjana Sultana Mim, for their invaluable guidance and support. They guided me like a kindhearted friend on whom I could rely on no matter what the situation. Their expert knowledge and vast experience have been a constant source of inspiration and motivation for me. I am grateful for their patience and encouragement, which have helped me overcome many challenges and achieve my goals, and have played a crucial role in the development of my dissertation.

I would like to express my deep appreciation to the members of my laboratory for their friendship and support. In particular, I would like to express my gratitude to Dr. Shun Kiyono for his invaluable assistance throughout my research. Additionally, I would like to thank Ms Haruka Aizawa, Mrs Mayumi Sugawara, and Mrs Yoriko Isobe for making my time in the laboratory comfortable and for their efforts in assisting me even with my broken Japanese language skills. I would also like to extend my thanks to Mrs Shiono Hiromi from the GSIS department for her diligent reminders about important dates and deadlines for MEXT Scholarship. Their constant efforts and valuable assistance have

greatly contributed to my enjoyable yet productive time in Japan and have made my time at the laboratory a truly rewarding experience.

Lastly, I would like to express my heartfelt thanks to my family and friends for their love, support and encouragement throughout my PhD journey. Their constant encouragement and belief in me have been a constant source of strength and motivation. Finally, I would like to dedicate this dissertation to my mother, who passed away before I had the opportunity to share this achievement with her. I know that she would have been so proud of me and I am deeply grateful for her love and support throughout my life.

I am deeply grateful to all of these wonderful people, and I sincerely hope that this dissertation serves as a small token of my appreciation for their invaluable support and guidance.

In Loving Memory of My Mother

Abstract

Automatically identifying implicit reasoning in arguments has been a challenging but important task that has recently gained significant attention in the computational argument mining community. The task involves formulating reasoning to explicate the implicit reasoning gap between argumentative components (i.e., claim and premise). Numerous studies in educational domain have shown that practicing such a task showed improved reasoning, logical, and argumentative skills in students. This provides us a strong motivation to develop an automatic implicit reasoning explication system that can be used in downstream applications, such as, assisting students write well-reasoned arguments, providing reasoning-based feedback or design models that can better understand human arguments.

Prior works have mainly relied on a supervised approach for explicating implicit reasoning in arguments, i.e., training/fine-tuning a generative model with manually written implicit reasonings as labels. These approaches rely on utilizing large language models (LLMs) for generation, that have been pre-trained on enormous amount of world-knowledge. While such LLMs generalise better than the previous state-of-the-art models, like RNNs (i.e., Recurrent Neural Networks), their generation quality still suffers to logically fill the implicit reasoning gap between a claim and its premise.

To overcome the aforementioned problem, in this thesis we propose a domain-specific approach towards implicit reasoning explication. Specifically, we hypothesize that (i) the reason why LLMs lack the ability to reasonablly explicate implicit reasoning in arguments might be due to the fact that they are trained on vast amount of domain-general knowledge, (ii) As an typical argument is usually based on a specific domain and its associated specific knowledge, incorporating small amount of such domain-specific knowledge might assist LLMs to enhance its reasoning. Towards this, we firstly develop a novel data creation methodology that can be used to create labeled domain-specific knowledgebase (KB). Secondly, we utilize this KB to train domain-specific models which then can be utilized for automatically explicating implicit reasonings.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> An argument is not won by shouting the loudest or by being the most aggressive. It is won by presenting the best evidences with reasoning.
>
> — *Excerpt from a Debate*

Arguments have become an essential part of our day-to-day communication. Using arguments to persuade one to accept a particular point of view or course of action, also referred to as argumentation, allows people to express their opinions or ideas, and to either convince others or defend themselves against opposing viewpoints such as in debates, online forums, classrooms, news, legal proceedings etc. Regardless of the way we argue, a typical argument comprises at-least two key components: (i) a clear and concise thesis statement i.e., the claim and, (ii) a supporting or undermining statement i.e., the premise, directed towards the claim through a well-reasoned logic.

It is widely argued that argumentation and reasoning often go hand-in-hand. For instance, to understand an argument one needs to logically reason through a premise in order to justify its claim. Such reasons or logical connections, referred to as implicit reasonings, are often unstated and are inferred by the listener or reader, acting as a bridge for correctly understanding the argument. In other words, one can reason through a premise towards its claim with the help of implicit reasonings. Since this process of comprehending an argument is crucial to how humans understand an argumentative discourse, it has become a widely researched theme in natural language processing (NLP) community.

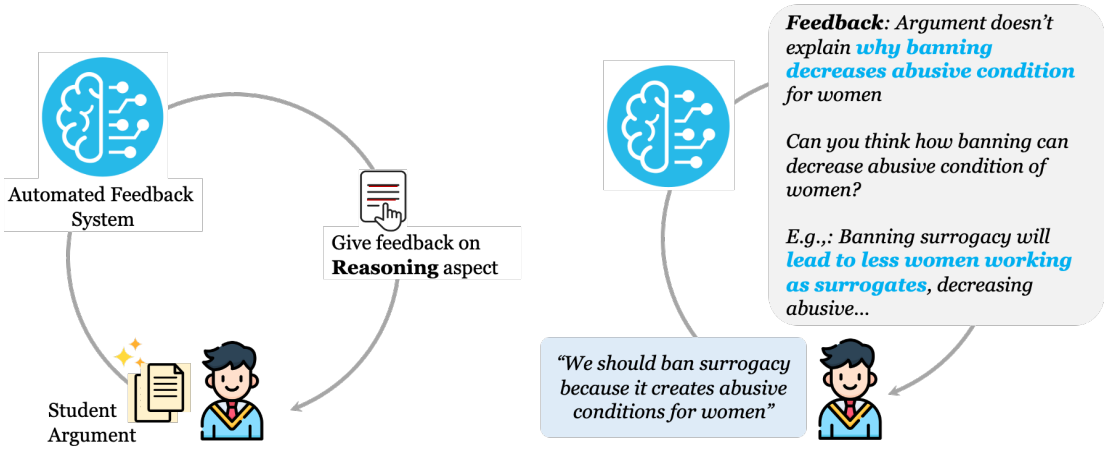

Figure 1.1: An automatic explication system can assist students to make well-reasoned arguments with automatic feedback

Argument mining, a sub-field of NLP, deals with automatically identifying, extracting, and analyzing argumentative components as well as reasoning in textual arguments. It has become an increasingly popular field within NLP due to its importance in developing downstream practical applications for law, education, politics, etc. A few examples of such applications include automated decision making, fact checking, evidence detection, argument search, argumentative writing support and feedback system, etc.

Argument mining is particularly useful in developing applications in educational domain due to its direct effect on students' comprehension skills. Specifically, research in educational domain has shown that students who practice identifying implicit reasonings in arguments develop better reasoning, argumentative writing and critical thinking skills. However, such practice require manual supervision by teachers which is an extremely time consuming and laborious task. Moreover, current automated writing support systems do not have the capability to identify implicit reasoning in arguments so as to help students write well-reasoned arguments.

In order to develop an automated writing and feedback support systems that focuses on reasoning aspect of arguments, as shown in Figure 1.1, in this work we address the task of automatic implicit reasoning explication which can be used to assist students write well-reasoned arguments and infer the hidden reasoning gap between claim and premise.

For explicating reasoning that is implicitly asserted in a given argument, knowing the implicit knowledge that makes part of such reasoning is crucial. Here, implicit knowledge refers our background knowledge that we use to logically reason from a premise

to its claim. Existing studies use either expert annotations of implicit reasonings or generative LLMs to explicate implicit knowledge and map them to a pre-defined syntactic structure of an implicit reasoning (Boltužić and Šnajder, 2016; Becker et al., 2017; Habernal et al., 2018a; Hulpus et al., 2019; Chakrabarty et al., 2021). However, for a given argument, the background knowledge gathered from LLMs is very generic which makes it difficult to frame implicit reasoning that sufficiently links claim and premise, and expert annotations tend to be very costly and lack scalability. In summary, LLMs lack knowledge specific to a particular topic (hereby referred to as domain) on which the argument is based. This is crucial to correctly frame the implicit reasoning that is specific to that argument. Moreover, manually crafting implicit reasonings by leveraging expert annotators can become a exceedingly costly process if more annotations are needed to train a generative model for automatically explicating implicit reasonings. To overcome the above challenges, in this thesis, we aim to automatically explicate implicit reasonings in arguments in a way that relies on knowledge specific to that domain, where such domain-specific knowledge can be gathered via a scalable yet less expensive annotation process.

In brief, in this thesis, we split our major task of automatically explicating implicit reasonings into two sub tasks. Firstly, we devise a annotation methodology to crowdsource domain-specific implicit reasonings at a reasonable cost and quality. Secondly, we propose a domain-specific approach towards automatic explication of implicit reasonings by fine-tuning generative LLMs on our crowdsourced domain-specific data.

1.1 Research Issues

In this thesis, we address the following research issues:

- **What is the appropriate methodology and format for annotation of domains-specific implicit reasonings?** Previous works have relied on either expert annotations or crowdsourcing to capture implicit reasoning between a given claim and premise. However, previous annotation methodologies are either too strict or too shallow to ensure quality as well as coverage of captured implicit reasonings. In our work, in order to overcome the above shortcomings, we take a middle approach and propose a semi-structured annotation framework for gathering implicit reasonings at a large-scale at reasonable cost and quality.

- **How to use the domain-specific knowledgebase of implicit reasonings for automatic explication?** Existing work have not yet explored the idea of using a

domain-specific knowledgebase for explicating implicit reasoning in argument. In our work, we take a domain-specific approach and assume that LLMs trained on small amount of domain-specific data can outperform the current state-of-the-art implicit reasoning explication models that rely on domain-general resources. We assume such a model to be analogous to student workbooks used to practice argumentation where topics/domain are known in advance.

- **What are the application of creating such a domain-specific resource?** Utilizing a domain-specific resource is comparatively new in argumentation domain and not well studied. We explore its application for a well-studied task, namely, evidence detection task i.e., given a claim and set of evidences, identify the most appropriate evidence that supports the claim.

1.2 Contributions

This thesis makes the following contributions:

- **Designing a semi-structured annotation framework to capture domain-specific implicit reasonings at scale:** We extensively analyze what representations (free-text form, semi-structured form, structured form) of implicit reasoning best enable us to capture the underlying logic between claim and premise. Based on our analysis, we design a novel semi-structured annotation framework that can be used with non-expert annotators and is suitable for large-scale crowdsourcing.

- **Construction of a domain-specific corpus using the proposed annotation framework:** We conduct multiple annotation studies and create 6 domain-specific corpus of implicit reasoning for wide variety of arguments for each domain. Our annotation study shows high coverage of our annotation framework for annotating implicit reasonings as well as high quality of crowdsourced annotations. Additionally, we show that with our methodology, creating such a corpus can be done overnight at a reasonable cost.

- **Establishing a domain-specific approach towards explicating implicit reasoning in arguments:** We test our hypothesis of taking a domain-specific approach by utilizing our domain-specific corpora and training generative models in a supervised approach. We show that our domain-specific model outperforms domain-general models even when using small number of training data. Additionally, we qualitatively evaluate the generation quality of implicit reasonings

(generated via domain-specific and domain-general models) and find our approach to result in significantly better structured and logical implicit reasonings.

- **Baseline model experiments for the automatic identification of reasoning patterns:** We consider the created domain-specific knowledgebase of implicit reasonings as an additional source for improving related argumentation task. Specifically, we leverage implicit reasonings for testing performance gain in Evidence Detection task. We treat it as a classification task, where, given claim and candidate evidence, the task is to classify evidence as acceptable or not. Our experiments show that while using domain-specific or domain-general implicit reasonings in a given classification model show similar results, they considerably improve upon current best performing LLMs such as BERT.

1.3 Thesis Overview

The rest of this thesis is structured as follows:

- **Chapter 2: Background**. In this chapter, first we introduce the basics of argumentation, its components and relations between them. Later we shed light on what is meant by implicit reasoning in relation to argumentation and how it is significantly important to bridge the implicit logical gaps between argumentative components to better understand arguments.

- **Chapter 3: Methodology to create domain-specific knowledgebase of implicit reasonings**. In this chapter, we present our analysis, annotation framework and crowdsourcing methodology designed to collect domain specific implicit reasonings. In brief, we first compare and analyze existing methodologies of annotation, and then propose a novel annotation design to overcome previous shortcomings. Later, we leverage our curated domain-specific knowledgebase for experimenting on automatic implicit reasoning explication task.

- **Chapter 4: Leveraging implicit reasoning knowledgebase for evidence detection task**. In this chapter, we test the applicability of our created resource(i.e., domain specific knowledgebase of implicit reasonings) for the downstream task of evidence detection. Specifically, we use implicit reasonings as an additional feature to classify candidate evidences given for a claim as acceptable or unacceptable. Later we present the results comparing domain-specific and domain-general approaches.

- **Chapter 5: Conclusion**. In the end we summarize our contributions and present our insights for future research.

Chapter 2

Background

This chapter introduces the basic notions related to arguments, argumentation and reasoning as well as gives a brief overview of argument mining and its downstream applications.

2.1 What is an Argument and Implicit Reasoning?

> *The first step toward clear thinking is the recognition that reasoning may be implicit as well as explicit*
>
> — Stephen Toulmin , *The uses of argument, Cambridge University Press (1958)*

An argument usually comprises a set of statements, usually presented in a logical manner, that aims to persuade or convince the listener or reader of the validity of a particular claim (Walton et al., 2008). As shown in Figure 2.1 below, the structure of a basic argument comprises of a claim (i.e., a statement that presents the position or belief that the argument is trying to support) and a premise (i.e., a statement that offers support to the claim in the form of statistics, expert testimony, or other forms that provide a basis for the argument), that are arranged in order to make a persuasive case.

In addition to the above components, an argument typically consists of a implicit reasoning (also termed as warrants) that is usually unstated or assumed by the reader or the listener (Toulmin, 1958; Freeman, 1992; Hitchcock, 2003). While the claim is the

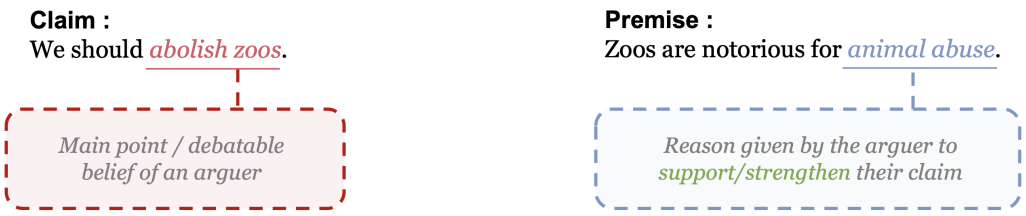

Figure 2.1: Basic components of an argument: (i) Claim and (ii) Premise

main point or position that is being argued for or against and premise is the information, facts, or examples that are used to support the claim, the implicit reasoning is the logical connection between the premise and the claim. It is the reasoning or explanation that connects the premise to the claim, and shows how the premise supports or justifies the claim. To better illustrate the function of an implicit reasoning, as shown in Figure 2.2, the implicit reasoning serves the purpose of linking the two argumentative components logically.

Although implicit reasonings help us better understand an argument and the relation between them, the process of deducing such a link happens relatively quickly in humans (National Academies of Sciences Engineering and Medicine and others, 2018), due to the vast amount of domain knowledge we possess as well as ability to reason with it (Hirschfeld and Gelman, 1994).

2.2 What is Argumentation?

Argumentation can be defined as the communicative activity of producing and exchanging reasons in order to support claims or defend/challenge positions, especially in situations of doubt or disagreement (Walton et al., 2008). In simple words, argumentation refers to using arguments to persuade, deliberate, convince or similar (Habernal et al., 2018b). For example, essays, debates, classroom discussions etc., are all forms of argumentation. However, most of the discourse we engage in are not instances of argumentation, for example when someone is having a query or when making a simple comment about an item. Argumentation is a process where an individual is asked to provide additional evidence or premises to support their claim. Essentially, argumentation serves as a crucial tool for discerning and evaluating information, rather than accepting information without questioning its validity. (Dutilh Novaes, 2022).

Argumentation plays an important role, especially in educational domain, where students can practice how to critically evaluate, convince and convey opinions via speech

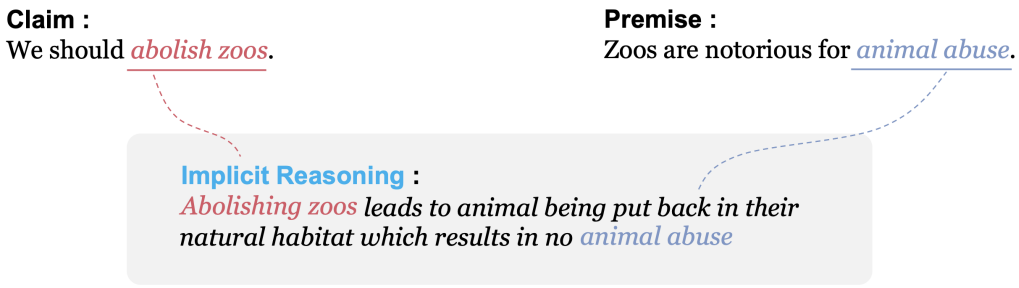

Figure 2.2: Understanding arguments usually involves using our knowledge about the world (i.e., Background Knowledge) to reason and bridge reasoning gap (implicit) between claim and premise.

or classroom discussions. Further, it has been shown that practicing such a discourse along with justifiable reasonings positively impacts their critical thinking, analysis and reasoning skills (Ennis, 1982; Erduran et al., 2004; von der Mühlen et al., 2019). However, since this kind of interactive learning requires student-teacher interaction which becomes a laborious task and lacks proper guidance and feedback when one teacher has to handle many queries by the students. Hence, recent approaches in Natural language processing have given rise to argumentation related studies for developing automatic systems capable of assisting teachers as well as students by giving feedback and writing assistance to each individual.

2.3 Argumentation Schemes

Argumentation schemes are standard patterns of templates for mapping different types of arguments in a well-defined structure. These schemes provide a framework for understanding how arguments are structured and how they can be evaluated. For instance, (Stab and Gurevych, 2014a) proposed their own scheme to model an argument into three components: Majorclaim, Claim and Premise, to annotate different components of an argument.

The most prominent and well known work in argumentation schemes is of (Walton et al., 2008), that proposes around 59 argumentation schemes to map wide-variety of arguments in a pre-defined template so as to better evaluate the strength and persuasiveness of arguments. An example from Walton's argumentation scheme is shown in Figure 2.3, where a given argument, when segmented into a claim and premise, can be used to understand the hidden assumptions like (i) Abolishing zoos has a consequence of no animal being abused; (ii) No animal being abused is a good consequence. In

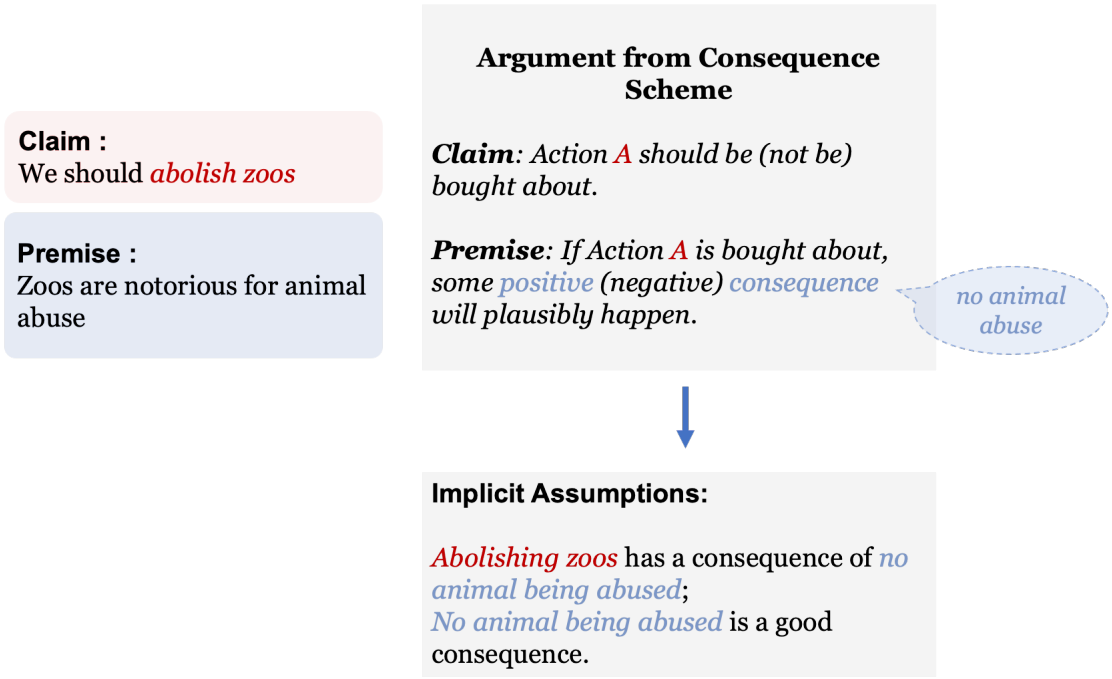

Figure 2.3: An example of Walton's Argument Schemes mapped on a typical argument

addition to the above schemes, (Toulmin, 1958) scheme extensively focuses on identifying one additional component i.e., the warrant or hidden assumptions in arguments. However, since such warrants are never explicitly stated, it becomes difficult to identify them.

2.4 Argument Mining and its Applications

Argument mining is a sub-field of NLP that has gained significant attention during the last decade. In brief, argument mining refers to the process of extracting and labeling argumentative components from text and organizing them in a structured format (Stab and Gurevych, 2014b). This process involves identifying the main claims, premises, and counterarguments within a text and organizing them in a way that makes them easy to understand and evaluate.

While argument mining is limited to argumentative text as main source of study, over the past few years, many challenging sub tasks been proposed that have multitude of potential downstream applications. At the higher-level, the task of identification of argumentative text from within a document was introduced as a starting point in argument mining (Palau and Moens, 2009; Reed, 2006; Peldszus and Stede, 2015; Kobbe et al., 2019). Later identification of much more fine-grained argumentative components like claim and premise was proposed (Stab and Gurevych, 2014b; Levy et al., 2014), fol-

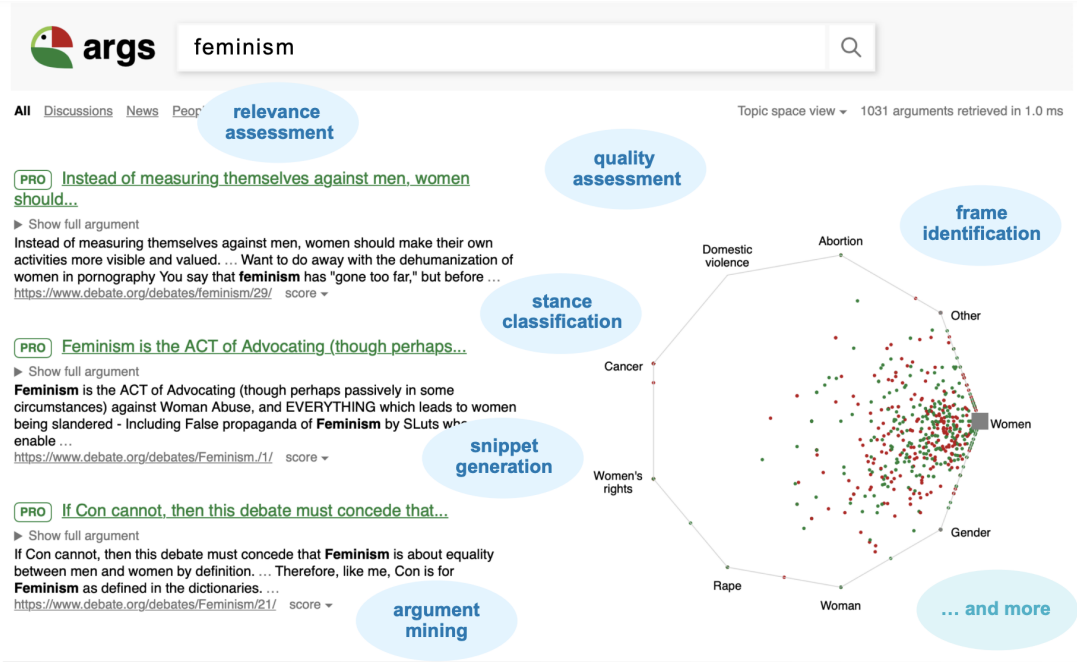

Figure 2.4: Wide range of argument mining related tasks and its application in developing downstream application. Args.me (Wachsmuth et al., 2017c) is one such application that helps find pro-con arguments for a given topic.

lowed by other explicit components like counter-arguments, evidences, facts (Rinott et al., 2015b; Aharoni et al., 2014b; Hua and Wang, 2017a,b; Reisert et al., 2018). With focus on argument structure and its application in essay scoring and quality evaluation, several works have proposed novel ways to correctly evaluate and analyze a given argumentative text (Persing et al., 2010; Persing and Ng, 2016; Wachsmuth et al., 2017a).

More recently, (Boltužić and Šnajder, 2016) proposed the task of automatically identifying implicitly asserted propositions in arguments which is closely connected to the study of arguments and implicit reasoning (Becker et al., 2017; Habernal et al., 2018a; Becker et al., 2020; Chakrabarty et al., 2021). Its important to note that identifying implicit argumentative components is still an open task and current models are not very capable in understanding the reasoning structure embedded within arguments (Becker et al., 2021).

Argumentation mining has also given rise to various downstream applications like argument search (Wachsmuth et al., 2017b), debating technology (Bar-Haim et al., 2021), fact checking (Samadi et al., 2016), automated decision making (Bench-Capon et al., 2009), writing support (Stab and Gurevych, 2017). More recently, constructive feedback assisted writing support (Wambsganss et al., 2020) was proposed to help learners/students improve their writing skills with the help of automatic feedback. However,

current feedback systems have basic functionality (e.g., suggest missing claims/premise or score argument based on pre-defined rubrics) and are not fully equipped to help the student learn why an argument is bad or how to make a well-reasoned argument.

Ongoing research into explicating the implicit reasoning gap in arguments (Singh et al., 2022; Becker et al., 2021) as well as how to give critical feedback in a writing assistant (Naito et al., 2022; Mim et al., 2022) might be the next step to help such systems more productive and useful for end application.

Chapter 3

Methodology to Create Domain-specific KnowledgeBase (KB) of Reasonings Overnight

3.1 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).

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. For example, consider the arguments comprising a claim and its premise, as shown in Fig. 3.1. 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 min-

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

Figure 3.1: Implicit Reasoning links the key-words in claim (labeled red) and premise (labeled blue) with unstated(background) knowledge.

ing 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 pre-trained 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 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, in this chapter we propose a crowdsourcing-based approach for collecting 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. 3.2), 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,

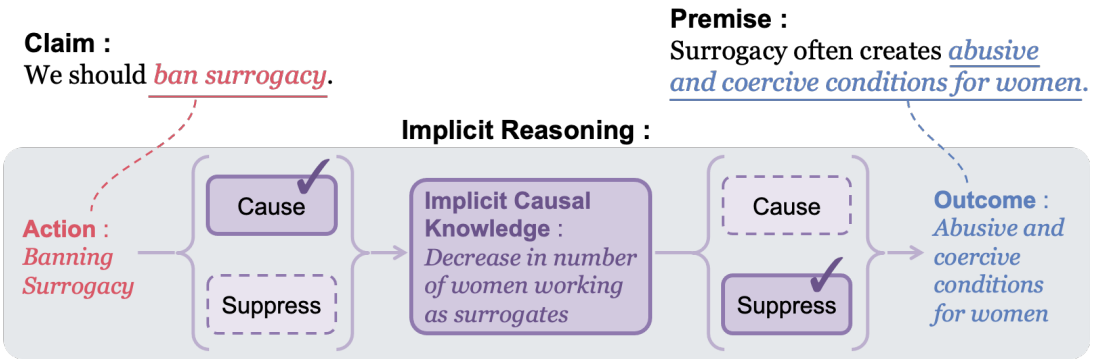

Figure 3.2: An example of our proposed semi-structured format to explicate implicit reasoning in arguments. Action and outcome represent the key-words/phrases derived from claim and premise respectively. The directed edges between action and outcome are causally linked via implicit causal knowledge, which explains the reasoning link between action and outcome.

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 chapter, 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.

3.2 Background

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 expensive to perform on a large scale. To overcome the prior challenges, (Habernal et al., 2018a) 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.

3.3 Semi-structured Implicit Reasonings

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. 3.2, 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. 3.2, 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. 3.2, 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. 3.2), 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.

Fig. 3.2 shows the final implicit reasoning representation in a semi-structured format along with the other aforementioned components.

3.4 Crowdsourcing Semi-structured Implicit Reasoning

We design a two-phase annotation process to obtain high-quality semi-structured implicit reasonings on a large scale (shown in Fig. 3.3), where each phase (§ 3.4.1 and § 3.4.2) 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. 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

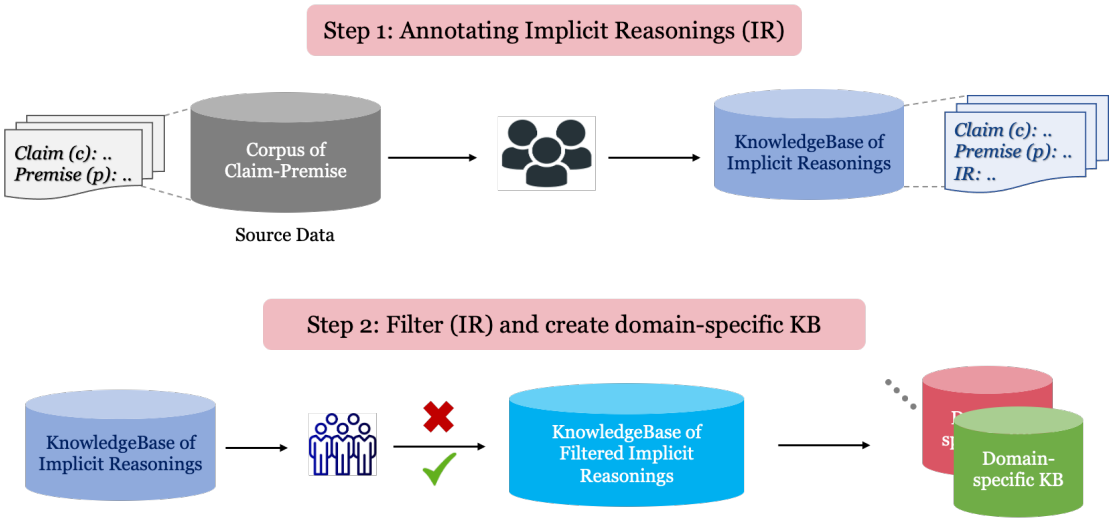

Figure 3.3: Overview of our methodology for creating domain-specific KB of Implicit Reasonings

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.

3.4.1 Phase 1: Framing Implicit Reasoning

In order to frame semi-structured implicit reasoning, we need four main components, 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*

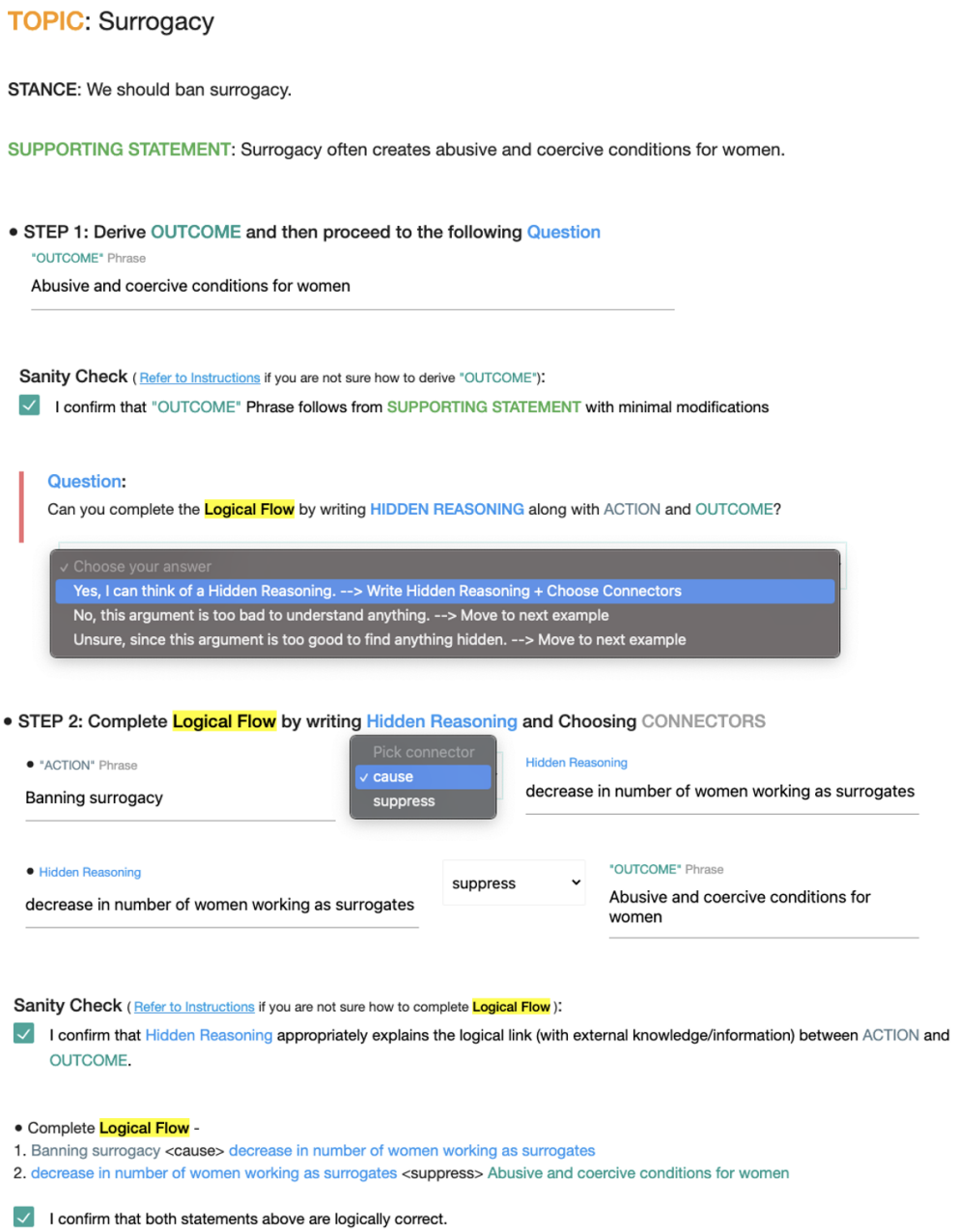

Figure 3.4: The interface of our crowdsourcing task for Phase 1. This phase consists of two steps, where STEP 1 is mandatory while STEP 2 depends on the choice made by crowdworkers for the Question preceding STEP 2.

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. 3.4, where in Step 1 annotators are asked to derive the outcome entity for a given premise [1].

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. 3.4). 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. 3.4, where annotators are provided with a predefined template for constructing the relationship between action entity, out-

[1] 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.

come 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. 3.4, 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.

3.4.2 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., 2017a), 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.

3.4.3 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.

Topic	# Claim-Premise	# IR	IRs ≥ 1	IRs ≥ 2	Avg. # IR per Premise
School uniform	145	483	99%	95%	3.3 (144)
Punishment	176	322	86%	60%	2.1 (152)
Zoos	141	390	98%	86%	2.8 (139)
Whaling	164	468	96%	83%	3.0 (158)
Voting	116	376	100%	94%	3.2 (116)
Cannabis	210	597	95%	86%	2.9 (200)
Total	952	2636	95%	83%	2.9 (909)

Table 3.1: 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.

3.5 IRAC dataset

3.5.1 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 3.1 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 3.1, 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

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

annotation from our final curated dataset is shown in Table 3.2, 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.

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 work-time 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

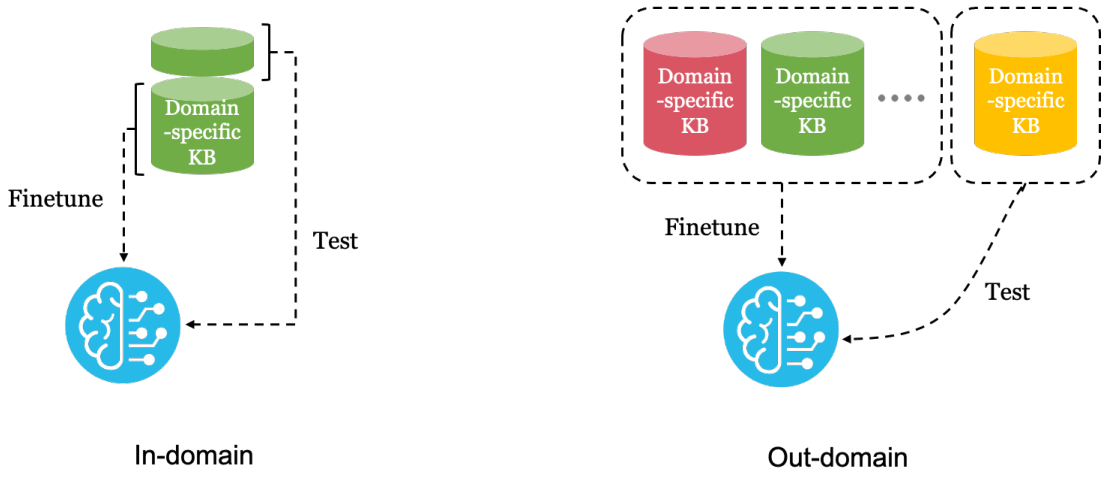

Figure 3.5: Two setting used in our experiments: In-domain (left) and Out-domain (right)

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.

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

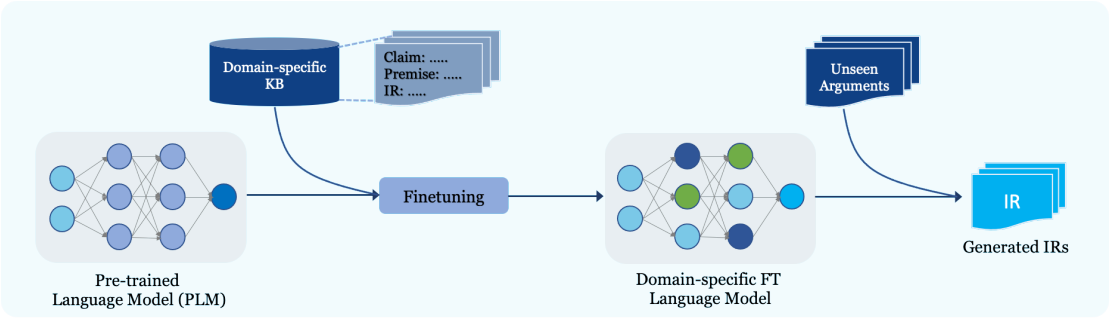

Figure 3.6: Overview of our approach for explicating implicit reasonings in in-domain setting.

3.6 Automatic Explication of Implicit Reasonings

3.6.1 Experiments

3.6.1.1 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.

3.6.2 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, as shown in Fig. 3.5, 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 fine-tune 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, as shown in Fig. 3.6, we fine-tune 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.

3.6.3 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 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*} *IR*. *And IR* {*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.

Domain	Baseline			Our Model		
	Bleu-1	Bleu-2	BertScore F1	Bleu-1	Bleu-2	BertScore F1
	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 3.3: Automatic evaluation of implicit reasoning (generation by fine-tuned BART) in two settings based on BLEU1 (B1), BLEU2 (B2) and BERTScore (BS).

3.6.4 Results and Analysis

As shown in Table 3.3, 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 3.4.

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.

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 3.4: 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.

3.7 Conclusion

In this chapter, we developed the conceptual foundation of this thesis. We investigated the role semi-structured implicit reasoning plays when bridging the gap between claim and premise. We also looked at the different approaches from which implicit reasonings can be annotated, comparing them regarding their degree of quality, and found that semi-structured implicit reasoning better link the reasoning gap between claim and premise.

Next, we applied out annotation framework to crowdsource implicit reasoning at scale. We created 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. 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. We will also study the effect of varying size of training data on the generating capability of our model.

Chapter 4

Expanding the Application of Domain-specific Implicit Reasonings for Improving Evidence Detection task

4.1 Introduction

An argument is composed of two key components: *claim*, i.e., a debatable belief or opinion, and *a supporting piece of statement*. Identification of these components and predicting the relationship among them forms the core of an important research area in argument mining (Peldszus and Stede, 2013) because they have become an essential component in building downstream natural language systems capable of arguing, debating, and fact checking (Rinott et al., 2015a; Lippi and Torroni, 2016; Alhindi et al., 2018; Lytos et al., 2019; Slonim et al., 2021).

Evidence detection (Aharoni et al., 2014a) is a sub-task in argument mining that has rose to prominence due to its direct relevance in building the aforementioned applications. Specifically, 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). In order to better illustrate evidence detection task, shown in Fig. 4.1 is an example of a given claim and three candidate evidences. In this example, identification of the best supporting piece of evidence is challenging as all three evidence are related to the claim, and only second candidate evidence is acceptable. The first candidate qualifies as a opinion or a premise

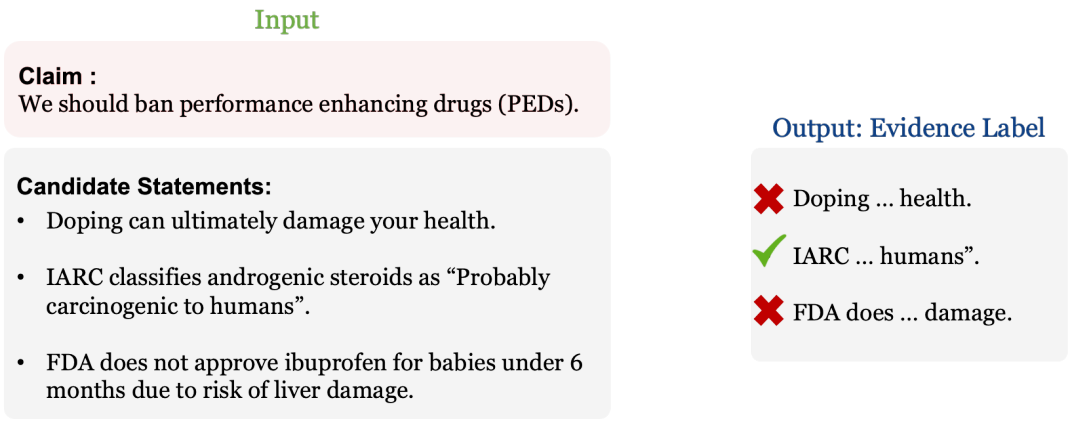

Figure 4.1: Three evidence candidate statements are given for a claim, where second candidate statement can be considered the best evidence piece.

that still supports the claim but does not contain any facts or judgements to back up the opinion, while candidate three, though being a factual statement or judgement and qualifies as an evidence, is irrelevant to PEDs and only mentions about the ill effects of drug use.

Towards solving the challenge of identifying acceptable evidence automatically, recent approaches have heavily relied on pretrained large language models (LLMs) as a default choice because of their outstanding performance in a wide range of NLP tasks (Howard and Ruder, 2018; Gururangan et al., 2020; Shnarch et al., 2018; Reimers et al., 2019; Elaraby and Litman, 2021). While most of these approaches use supervised learning (i.e., incorporating labeled data for training) and rely on the better generalization ability of LLMs (Devlin et al., 2018; Yang et al., 2019), 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 (Stahlhut, 2019; Sun et al., 2019).

Other approaches for evidence detection have also tried using lexical features extracted from argument components such as semantic similarity, adjacent sentence relation and discourse indicators (Stab and Gurevych, 2014b; Rinott et al., 2015b; Nguyen and Litman, 2016; Hua and Wang, 2017a). However, no prior work has considered identifying the underlying, implicit reasoning, also referred to as *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 an implicit reasoning between the claim and a piece of evidence as shown in Fig. 4.1, for candidate statement 2, the most plausible evidence piece could be detected as it has an underlying implicit link that can be established with the claim (i.e., *Drugs that are carcinogenic to humans should be banned*). In this chapter, we hypothesize that for detecting the best piece of evidence for a claim,

it is crucial to capture such implicit reasoning between them (Habernal et al., 2018b).

In order to validate our assumption and improve current evidence detection systems, 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., training 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 (Singh et al., 2019). 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.

4.2 Proposed Method

4.2.1 Overview

Given a query 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 § 4.1, 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. implicit reasonings). Our framework first extracts multiple implicit reasonings that link a given claim to an evidence piece, and later leverages the acquired implicit reasonings to estimate the likelihood of the claim being supported by that candidate evidence.

As described in § 4.1, we take a closed-domain approach (i.e., we train and test one

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

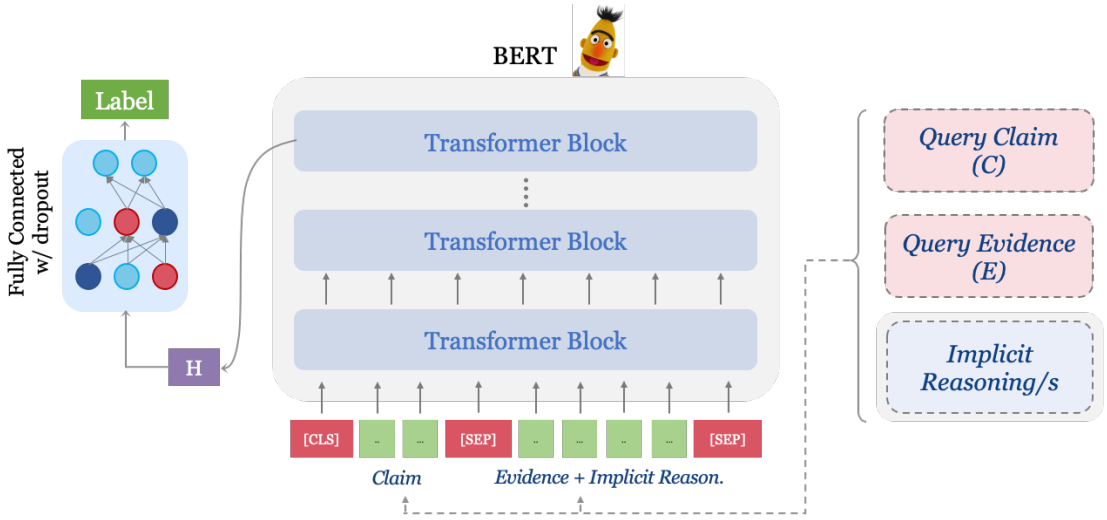

Figure 4.2: Overall framework for evidence detection task. We use BERT classifier with claim, evidence and extracted implicit reasoning as input features

topic at a time) and simultaneously leverage domain-specific implicit reasonings that are extracted via the implicit reasoning component (See § 4.2.2). The complete overview of our evidence detection framework is shown in Fig. 4.2.

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.

4.2.2 Implicit Reasonings Component

Extracting Implicit Reasonings 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 $\mathcal{D} = \{(c_i, p_i, r_i)\}_{i=1}^{n}$ 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

[2]In this work, the utilized source datasets $\mathcal{D}$ of implicit reasonings consists of premise instead of evidence. For more details, refer to Habernal et al. (2017); Singh et al. (2022)

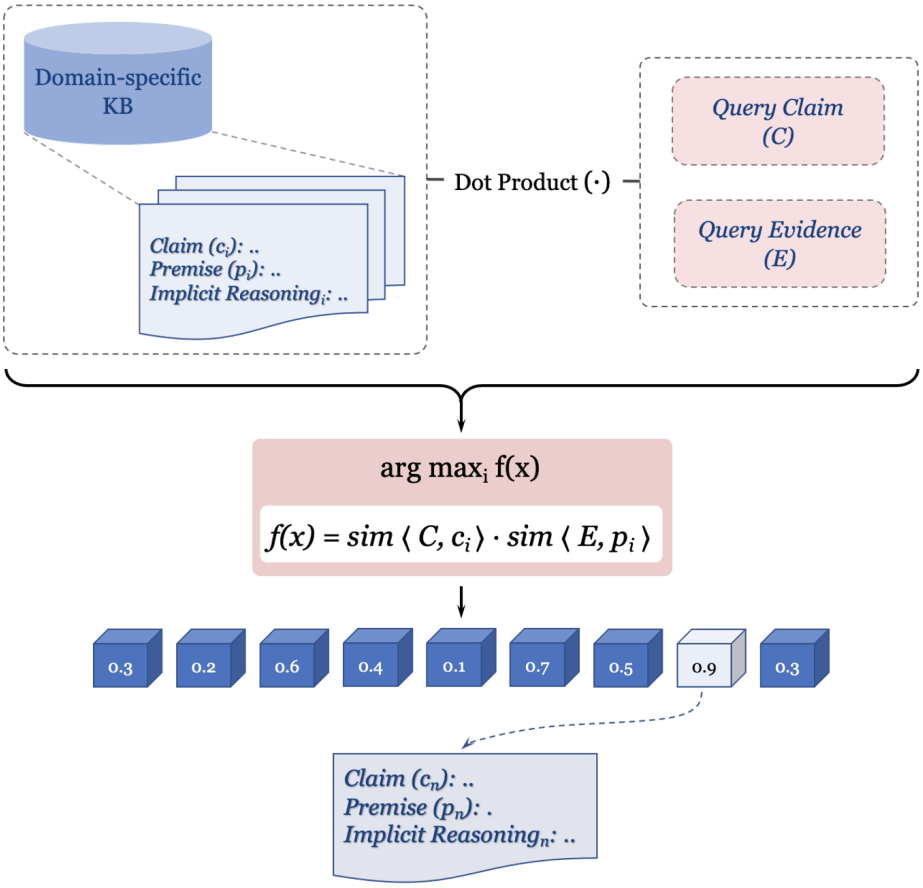

Figure 4.3: Overview of our methodology for extraction of reasonings from domain-specific knowledgebase

$\mathcal{D}$, as shown in Fig. 4.3. Specifically, we retrieve the top-m most 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 (Reimers and Gurevych, 2019), a BERT (Devlin et al., 2018) 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.

4.3 Experiments

4.3.1 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 Causal-

Topic	Acceptable Evidence (A)	Unacceptable Evidence (U)	Total	Ratio (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

Figure 4.4: We use IBM dataset as our source for claim-evidence instances

ity) (Singh et al., 2022), 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) (Habernal et al., 2017), 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. We utilize the dataset of the Argument Reasoning Comprehension Task (ARCT) (Habernal et al., 2018b), because it provides a large collection of implicit reasonings-annotated arguments that cover a wide variety of topics.

Evidence Data Instead of creating a dataset of claim and evidence pairs from nothing, we utilize the IBM-Evidence dataset (Ein-Dor et al., 2020). 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

[3]In the original paper, Habernal et al. (Habernal et al., 2017) refers to implicit reasonings as warrants.

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 Fig. 4.4.

4.3.2 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.

4.3.3 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 (Devlin et al., 2018), namely $\textbf{BERT}_{\textbf{base}}$, which has been shown to outperform the previously established state-of-the-art on similar tasks (Reimers et al., 2019; Thorne et al., 2018; Stahlhut, 2019). (ii) & (iii) Two separate models to additionally consider the implicit reasonings available via domain-specific or domain-general resource, namely $\textbf{BERT}_{\textbf{in}}$, and $\textbf{BERT}_{\textbf{out}}$ respectively. (iv) Additionally, we consider a random baseline that predicts the most frequent class label as observed in the training data.

4.3.4 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 Fig. 4.4), 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.

Random : Predicts most frequent class label as observed in training data

BERT$_{base}$: Baseline model **without** implicit reasonings

BERT$_{out}$: Model with **domain-general** implicit reasonings

BERT$_{in}$: Proposed model with **domain-specific** implicit reasonings

Evaluation Metrics* :
Precision (P)
Recall (R)
F1-Score (F1)

*Macro-averaged

Domain	**Random**			**BERT$_{base}$**			**BERT$_{in}$**			**BERT$_{out}$**		
	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

Figure 4.5: Classification accuracy of BERT in domain-specific setting. We experiment with different variations of BERT i.e., with and without implicit reasonings.

4.3.5 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 Fig. 4.5, all BERT-based models beat the random baseline on all topics, except *Ban whaling*, where their performance is marginally higher. **BERT$_{in}$** outperforms **BERT$_{out}$** in all topics except *Ban whaling* and *Compulsory voting*. Overall, **BERT$_{base}$** 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$_{in}$** achieved higher performance for only two topics.

4.3.6 Qualitative Analysis of Implicit Reasonings

Contrary to our expectation, **BERT$_{base}$** 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$_{out}$** performance decreased for

Claim:	**W**e should ban zoos
Candidate Evidence (acceptable):	**M**any animal rights activists argue that zoo animals often suffer due to the transition from being free and wild to captivity
Domain-specific Implicit Reasoning:	**A**bolishing zoos suppresses caging animals who run free and caging animals who run free causes animals suffer from mental abuse
Domain-general Implicit Reasoning:	**H**aving too many visitors at once ruins the experience for everybody

Figure 4.6: Example of implicit reasonings extracted for a given query claim-evidence

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.

An example of an instance from IBM evidence dataset along with our extracted implicit reasoning is shown in Fig. 4.6. The domain-specific implicit reasoning that is extracted from our domain-specific knowledgebase i.e., IRAC (Singh et al., 2022), explains the causal reasoning between *"abolishing zoos"* and *"animal suffering from mental abuse"*, that might help the the classification model to correctly identify the acceptable evidence. One the other hand, the domain-general implicit reasoning from ARCT dataset (Habernal et al., 2018a) does not add any reasonable link to the given claim and candidate evidence.

To further investigate how well implicit reasonings help in evidence detection task, we ask two expert annotators to judge 50 randomly selected instances along with the relevant implicit reasonings on the following questions: (i) Do implicit reasoning help deduce acceptability of evidence? and, (ii) Which implicit reasoning is more relevant to a given claim-acceptable evidence? For (i) On an average, 32 out of 50 instances were answered "yes" by both annotators with a krippendorff's α of 0.72, depicting that implicit reasonings do assist even humans to judge the acceptability of an evidence. Note that, for (i) we provided implicit reasonings from both domain-specific as well as domain-general knowledgebase and the annotators could chose either one of them or none. For (ii) only 28 out of 50 instances were marked in favor of domain-specific implicit reasoning, while only 8 were marked for domain-general reasonings. Overall krippendorff's α was 0.58, depicting moderate agreement. This shows that while domain-specific implicit reasoning were better than domain-general reasonings, their

relevancy to a given claim and evidence might not always be acceptable. We assume this to be due to the extraction method we use might not be sufficient and can be improved further.

4.4 Conclusion and Future Work

In this chapter, 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.

Chapter 5

Conclusion

In conclusion, the research presented in this thesis has proposed an ”domain-specific approach” for bridging the implicit reasoning gap in arguments. This approach utilizes a combination of natural language processing and machine learning techniques to identify and generate implicit reasoning for a given piece of argument. The results of our experiments demonstrate that this approach is effective at uncovering implicit reasoning, and can be applied to a wide range of argumentative texts.

The proposed approach has the potential to improve the way we analyze and evaluate arguments by providing a more complete picture of the reasoning behind them. It can also aid in the development of more sophisticated AI systems for argumentation and decision-making. Furthermore, it can be used as a tool for helping people to improve their critical thinking skills by identifying implicit reasoning in their own arguments and those of others.

Additionally, in this thesis we explored the utilization of implicit reasonings for improving evidence detection task. We showed that by leveraging the ability of implicit reasonings to infer hidden connections and relationships within data, we can improve the accuracy and efficiency of evidence detection, resulting in more effective and efficient decision-making by our model.

In future work, we plan to improve the efficiency of the model and test it on a larger dataset to further establish its effectiveness. Additionally, we will explore applications of the model beyond argumentation, such as in the areas of legal reasoning and political discourse. We believe that this approach has the potential to make a meaningful contri-

bution to the field of argumentation and artificial intelligence, and we look forward to seeing its continued development and impact.

References

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

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 2014a. Association for Computational Linguistics. doi: 10.3115/v1/W14-2109. URL https://aclanthology.org/W14-2109.

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, 2014b.

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. End-to-end argumentation knowledge graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7367–7374, 2020.

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.

Roy Bar-Haim, Liat Ein-Dor, Matan Orbach, Elad Venezian, and Noam Slonim. Advances in debating technologies: Building AI that can debate humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–5, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-tutorials.1. URL https://aclanthology.org/2021.acl-tutorials.1.

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

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

Maria Becker, Siting Liang, and Anette Frank. 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 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.deelio-1.2. URL https://aclanthology.org/2021.deelio-1.2.

Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. Altruism and agents: an argumentation based approach to designing agent decision mechanisms. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1073–1080. Citeseer, 2009.

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.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. Implicit premise generation with discourse-aware commonsense knowledge models. *arXiv preprint arXiv:2109.05358*, 2021.

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.

Catarina Dutilh Novaes. Argument and Argumentation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

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.

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.

Robert H Ennis. Identifying implicit assumptions. *Synthese*, pages 61–86, 1982.

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

Vanessa Wei Feng and Graeme Hirst. 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 2011. Association for Computational Linguistics.

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

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

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.

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.

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*, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 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, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1175. URL https://www.aclweb.org/anthology/N18-1175.

Lawrence A Hirschfeld and Susan A Gelman. *Mapping the mind*. Citeseer, 1994.

David Hitchcock. Toulmin's warrants. In *Anyone who has a view*, pages 69–82. Springer, 2003.

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

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

Xinyu Hua and Lu Wang. 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, July 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-2032. URL https://www.aclweb.org/anthology/P17-2032.

Xinyu Hua and Lu Wang. Understanding and detecting supporting arguments of diverse types. *arXiv preprint arXiv:1705.00045*, 2017b.

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

Ralph Henry Johnson and J Anthony Blair. *Logical self-defense*. Idea, 2006.

Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. Exploiting background knowledge for argumentative relation classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

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

John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, December 2019. doi: 10.1162/coli_a_00364. URL https://aclanthology.org/J19-4006.

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, 2014.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

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.

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.

Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, Keshav Singh, and Kentaro Inui. Lpattack: A feasible annotation scheme for capturing logic pattern of attacks in arguments. *arXiv preprint arXiv:2204.01512*, 2022.

Shoichi Naito, Shintaro Sawada, Nakagawa Chihiro, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh, and Kentaro Inui. Typic: A corpus of template-based diagnostic comments on argumentation. *arXiv preprint*, 2022.

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

Huy Nguyen and Diane Litman. 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, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1107. URL https://www.aclweb.org/anthology/P16-1107.

Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, January 2013. ISSN 1557-3958. doi: 10.4018/jcini.2013010101. URL http://dx.doi.org/10.4018/jcini.2013010101.

Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815, 2015.

Isaac Persing and Vincent Ng. Modeling stance in student essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2174–2184, 2016.

Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, 2010.

Chris Reed. Preliminary results from an argument corpus. *Linguistics in the twenty-first century*, pages 185–196, 2006.

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. URL http://arxiv.org/abs/1908.10084.

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. doi: 10.18653/v1/P19-1054. URL https://aclanthology.org/P19-1054.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 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 2018.

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 2015a. Association for Computational Linguistics. doi: 10.18653/v1/D15-1050. URL https://aclanthology.org/D15-1050.

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 2015b. Association for Computational Linguistics. doi: 10.18653/v1/D15-1050. URL https://www.aclweb.org/anthology/D15-1050.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. Explagraphs: An explanation graph generation task for structured commonsense reasoning. *arXiv preprint arXiv:2104.07644*, 2021.

Mehdi Samadi, Partha Talukdar, Manuela Veloso, and Manuel Blum. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

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.

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. doi: 10.18653/v1/D19-6610. URL https://aclanthology.org/D19-6610.

Keshav Singh, Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, and Kentaro Inui. 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 2021. Association for Computational Linguistics.

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. URL https://aclanthology.org/2022.lrec-1.499.

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.

Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510, 2014a.

Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014b.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017.

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.

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.

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. doi: 10.18653/v1/W18-5501. URL https://aclanthology.org/W18-5501.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003. doi: 10.1017/CBO9780511840005.

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

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

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 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, 2017a.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, 2017b.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web. In Kevin Ashley, Claire Cardie, Nancy Green, Iryna Gurevych, Ivan Habernal, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker, editors, *4th Workshop on Argument Mining (ArgMining 2017) at EMNLP*, pages 49–59. Association for Computational Linguistics, September 2017c. URL https://www.aclweb.org/anthology/W17-5106.

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

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. A corpus for argumentative writing support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.74. URL https://aclanthology.org/2020.coling-main.74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

List of Publications

International Conference Papers (Refereed)

1. 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 13th International Conference on the Language Resources and Evaluation Conference (LREC), June 2022.

2. Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, Keshav Singh and Kentaro Inui. LPAttack: A Feasible Annotation Scheme for Capturing Logic Pattern of Attacks in Arguments. In Proceedings of the 13th International Conference on the Language Resources and Evaluation Conference (LREC), June 2022.

3. Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh and Kentaro Inui. TYPIC: A Corpus of Template-Based Diagnostic Comments on Argumentation. In Proceedings of the 13th International Conference on the Language Resources and Evaluation Conference (LREC), June 2022.

4. Keshav Singh, Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, Kentaro Inui. Exploring Methodologies for Collecting High-Quality Implicit Reasoning in Arguments. Proceedings of the 8th Workshop on Argument Mining, pages 57–66, November 10–11, 2021.

5. Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naitoh and Kentaro Inui. Annotating Implicit Reasoning in Arguments with Causal Links. Argumentation Knowledge Graphs, AKBC October 2021 (Online).

6. 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, November 3, 2019.

7. Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert and Kentaro Inui. When Choosing Plausible Alternatives, Clever Hans can be Clever. In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing (COIN), pages 33–42 Hongkong, China, November 3, 2019.

Other Publications (Not refereed)

1. Keshav Singh, Paul Reisert, Naoya Inoue, Kentaro Inui. Towards Understanding Implicit Reasoning in Arguments via Multiple Warrants 言語処理学会第27回年次大会 , pp.371-374, March 2021

2. Keshav Singh, Naoya Inoue, Paul Reisert and Kentaro Inui. Ranking warrants with pairwise preference learning. 言語処理学会第26回年次大会 , pp.776-779, March 2020

3. Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, Kentaro Inui. Balanced COPA: Countering Superficial Cues in Causal Reasoning. 言語処理学会第26回年次大会 , pp.1105-1108, March 2020

4. Keshav Singh, Edwin Simpson, Paul Reisert, Iryna Gurevych, Kentaro Inui. Improving Evidence Detection using Warrants as External Knowledge. 言語処理学会第25回年次大会 , pp.1241-1244, March 2019