

Automatic Generation of Structured Explanations for Arguments from Consequences

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Jonathan Lukas Kobbe
aus Zweibrücken

Mannheim, 2023

Dekan: Prof. Dr. Claus Hertling, Universität Mannheim
Referent: Dr. I.R. Ioana Karnstedt-Hulpuş, Utrecht University
Korreferent: Prof. Dr. Heiner Stuckenschmidt, Universität Mannheim

Tag der mündlichen Prüfung: 21.11.2023

Abstract

Argumentation is a complex means of communication which has been studied since many centuries. While argumentation generally aims to convince, it is also highly dependent of the context. Thus, many works on analyzing or mining arguments focus on a particular domain. In this thesis, we take a different perspective by addressing one particular type of arguments, called *arguments from consequences*. Our aim is to understand and automatically explain such arguments.

General approaches for explaining texts are an important step towards explaining arguments, but they most often lack to complete the coherence of arguments. Instead, we propose methods to specifically explain arguments from consequences in a formalized and well-defined way which facilitates downstream tasks such as (counter-) argument generation or large scale analyses of debates.

Our first step includes automatically finding modular representations for arguments from consequences. For evaluating these, we use them to detect the argument's stance (in favor / against). We expand upon related work in stance detection by proposing an unsupervised method which specifically addresses one particular type of arguments, but offers the advantage of being topic independent and explainable.

Further, in order to explain why the postulated consequence of an argument from consequences holds, we propose a method to extract *effect relations* from text. In contrast to related work, our proposed method is conceptually simple and does not involve training, but still achieves comparable results. We use this method on argumentative and, other than in related work, encyclopedic texts to generate a knowledge graph. Our evaluation shows that the graph has relatively high precision, but a low recall. Compared to related work, our graph is more comprehensive and publicly available. We use the graph to explain arguments from consequences by exploiting the transitivity of effect relations and evaluate the generated explanations a posteriori.

Lastly, for explaining why the consequence is considered good or bad, we propose to use the *moral foundations theory*. The classification of moral foundations is being

researched actively. We also train classifiers and, in contrast to related work, concretely evaluate their use on arguments. Further, we explore the usage of moral foundations in argumentative texts. We find their usage to be weakly correlated with argument quality and audience approval.

Summarizing, we propose methods to analyze and explain specifically arguments from consequences. This focus allows us to access the arguments' underlying reasoning, which we consider to be an important step towards the modeling of arguments and relevant background knowledge.

Zusammenfassung

Argumentation ist ein komplexes Kommunikationsmittel, das seit vielen Jahrhunderten erforscht wird. Während Argumentation im Allgemeinen darauf abzielt zu überzeugen, ist sie auch stark kontextabhängig. Daher konzentrieren sich viele Untersuchungen zur Analyse oder Gewinnung von Argumenten auf einen bestimmten Bereich. In dieser Arbeit nehmen wir eine andere Perspektive ein, indem wir uns mit einer bestimmten Art von Argumenten befassen, die als *Argumenta ad Consequentiam* bezeichnet werden. Unser Ziel ist es, solche Argumente zu verstehen und automatisch zu erklären.

Allgemeine Ansätze zur Erklärung von Texten stellen einen wichtigen Schritt zur Erklärung von Argumenten dar, sie reichen jedoch meist nicht aus, um die Kohärenz der Argumente zu vervollständigen. Stattdessen schlagen wir Methoden vor, um spezifisch *Argumenta ad Consequentiam* auf formalisierte und klar definierte Weise zu erklären, was nachgelagerte Aufgaben wie die Generierung von (Gegen-)Argumenten oder umfangreiche Analysen von Debatten erleichtert.

Unser erster Schritt besteht darin, automatisch modulare Darstellungen für *Argumenta ad Consequentiam* zu finden. Um diese zu evaluieren, nutzen wir sie, um die Haltung des Arguments (pro / kontra) zu erkennen. Wir erweitern verwandte Arbeiten zur Erkennung von Haltungen, indem wir eine unüberwachte Methode vorschlagen, die speziell auf eine bestimmte Art von Argumenten eingeht, aber den Vorteil bietet, themenunabhängig und erklärbar zu sein.

Um zu erklären, warum die postulierte Konsequenz eines Argumentum ad Consequentiam zutrifft, schlagen wir außerdem eine Methode vor, um *Effektrelationen* aus Text zu extrahieren. Im Gegensatz zu verwandten Arbeiten ist unsere vorgeschlagene Methode konzeptionell einfach und erfordert kein Training, erzielt aber dennoch vergleichbare Ergebnisse. Wir verwenden diese Methode auf argumentative und, anders als in verwandten Arbeiten, enzyklopädische Texte, um einen Wissensgraphen zu erstellen. Unsere Evaluation zeigt, dass der Graph eine relativ hohe Genauigkeit, aber eine geringe Sensitivität aufweist. Im Vergleich zu verwandten Arbeiten ist unser Graph

umfassender und öffentlich zugänglich. Wir verwenden den Graphen, um Argumenta ad Consequentiam zu erklären, indem wir die Transitivität der Effektrelationen ausnutzen, und evaluieren die generierten Erklärungen a posteriori.

Um schließlich zu erklären, warum die Konsequenz als gut oder schlecht angesehen wird, schlagen wir vor, die *Moral Foundations Theory* (Theorie moralischer Grundlagen) zu verwenden. Die Klassifizierung moralischer Grundlagen wird aktiv erforscht. Auch wir trainieren Klassifikatoren und evaluieren deren Verwendung, im Gegensatz zu verwandten Arbeiten, konkret auf Argumenten. Darüber hinaus untersuchen wir die Verwendung von moralischen Grundlagen in argumentativen Texten. Wir stellen fest, dass ihre Verwendung schwach mit der Qualität der Argumente und der Zustimmung des Publikums korreliert.

Zusammenfassend schlagen wir Methoden vor, um Argumenta ad Consequentia gezielt zu analysieren und zu erklären. Dieser Fokus ermöglicht uns den Zugang zu den Schlussfolgerungen, die den Argumenten zugrunde liegen, was wir als wichtigen Schritt zur Modellierung von Argumenten und relevantem Hintergrundwissen betrachten.

Contents

List of Publications	x
List of Acronyms	xii
1 Introduction	1
1.1 Problem Definition	4
1.2 Contributions	7
1.3 Outline	8
2 Theoretical Background and Related Work	9
2.1 Theoretical Background	9
2.1.1 Argumentation	9
2.1.2 Natural Language Processing	12
2.1.3 Knowledge Graphs	14
2.1.4 Moral Foundations Theory	16
2.2 Related Work	18
2.2.1 Explaining Arguments	18
2.2.2 Stance Detection and related tasks	21
2.2.3 Relation Extraction and Knowledge Graph Generation	26
2.2.4 Moral Foundation Classification	28
2.2.5 Contributions	29
3 Preliminary Work: External Knowledge for Argumentative Relation Classification	31
3.1 Knowledge Graph Features	33
3.2 Neural Network Model	35
3.3 Experiments	36
3.3.1 Data	36
3.3.2 Knowledge Graphs	38
3.3.3 Baseline	38
3.3.4 NN Model Optimization and Configurations	39
3.3.5 Results	39
3.3.6 Analysis	40
3.4 Conclusion	41

4	Analyzing Arguments from Consequences	43
4.1	Reconstructing the Premises: The Effect Triple	45
4.1.1	Lexicons	46
4.1.2	Effect Triple Extraction	47
4.2	Stance Detection with Effect Triples	49
4.3	Dataset Generation	52
4.3.1	Data Collection	52
4.3.2	Crowdsourcing Study	52
4.3.3	Agreement and Reliability	52
4.3.4	Final Dataset	54
4.4	Evaluation	55
4.4.1	Data	55
4.4.2	Compared systems	55
4.4.3	Results and Discussion	56
4.4.4	Error Analysis	58
4.5	Transformers for Stance Detection	59
4.5.1	In-Domain Evaluation	60
4.5.2	Cross-Domain Evaluation	60
4.5.3	Error Analysis	61
4.6	Conclusion	62
5	Explaining the Effect Premise: The Effect Graph	64
5.1	Effect Graph Generation	66
5.1.1	Effect Relation Extraction (EREx)	67
5.1.2	Graph Construction	67
5.2	Explanation Generation	69
5.2.1	Effect-Effect-Explanation	69
5.2.2	Effect-Lexical-Explanation	70
5.2.3	Explanation Candidate Filtering	71
5.3	Evaluation	72
5.3.1	Effect Relation Extraction Evaluation	73
5.3.2	Effect Graph Evaluation	74
5.3.3	Explanation Evaluation	83
5.4	Discussion	87
6	Explaining the Judgment Premise: Moral Foundations	89
6.1	Predicting Moral Sentiment in Tweets and Debates	91
6.1.1	A New Test Set for Moral Framing in Argumentation	91
6.1.2	Methods for the Prediction of Moral Sentiment	92
6.1.3	State of the Art and Baselines	94
6.1.4	Data	95
6.1.5	Results for MF Prediction on Tweets and Debates	96
6.2	Correlation Studies	97
6.2.1	Data	97

<i>CONTENTS</i>	ix
6.2.2 Results for the Correlation Analysis	98
6.3 Discussion	100
7 Conclusion	102
Acknowledgments	104
Bibliography	105
A Published Resources	121
B Relevant Dependency Relations	122
C Effect Relation Validity Annotation	123
D Explanation Quality Annotation	126

List of Publications

Most research presented in this thesis has been previously published. This includes text, figures, and tables. We reference these publications in the respective chapters. In the following, we list all of our publications which are related to this thesis:

- Kobbe, J., Hulpuş, I., and Stuckenschmidt, H. (2020a). Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.
- Kobbe, J., Opitz, J., Becker, M., Hulpuş, I., Stuckenschmidt, H., and Frank, A. (2019). Exploiting Background Knowledge for Argumentative Relation Classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASICS)*, pages 8:1–8:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.
- Kobbe, J., Hulpuş, I., and Stuckenschmidt, H. (2023). Effect graph: Effect relation extraction for explanation generation. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 116–127, Toronto, Canada. Association for Computational Linguistics.
- Kobbe, J., Rehbein, I., Hulpuş, I., and Stuckenschmidt, H. (2020b). Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.
- Hulpuş, I., Kobbe, J., Stuckenschmidt, H., and Hirst, G. (2020). Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online). Association for Computational Linguistics.
- Hulpuş, I., Kobbe, J., Meilicke, C., Stuckenschmidt, H., Becker, M., Opitz, J., Nastase, V., and Frank, A. (2019). Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ ISWC*, pages 62–77.
- Paul, D., Opitz, J., Becker, M., Kobbe, J., Hirst, G., and Frank, A. (2020). Argumentative relation classification with background knowledge. In *Computational Models of Argument*, pages 319–330. IOS Press.

- Becker, M., Hulpuş, I., Opitz, J., Paul, D., Kobbe, J., Stuckenschmidt, H., and Frank, A. (2020). Explaining arguments with background knowledge: Towards knowledge-based argumentation analysis. *Datenbank-Spektrum*, 20:131–141.

List of Acronyms

- AKG** Al-Khatib’s Argumentation Knowledge Graph. 74, 75
- BERT** Bidirectional Encoder Representation from Transformers. 14, 22, 56, 57, 59–63, 94, 96–98
- Bi-LSTM** Bidirectional Long Short-Term Memory. 35
- CN** ConceptNet. 15, 16, 32–34, 38–42, 87, 101, 102
- CORPS** Corpus of tagged Political Speeches. 97–100
- DB** DBpedia. 15, 16, 38–40, 93, 101
- DeBERTa** Decoding-enhanced BERT with Disentangled Attention. 14, 59–62
- dir** Direction. 45, 46, 48–51
- ECF** Extended Connotation Frames lexicon. 47, 55–57, 59, 67, 73
- eff** Effect. 45, 46, 49–51
- EREx** Effect Relation Extractor. 26, 27, 29, 66, 67, 69, 73–76, 80, 82–84
- EWN** +/-EffectWordNet. 47, 55–57, 59
- HIT** Human Intelligence Task. 52, 76
- IAA** Inter-Annotator Agreement. 20, 21, 52, 73, 78, 80, 87, 91, 95, 100, 101, 103
- KG** Knowledge Graph. 3, 8, 9, 14–16, 18, 21, 26, 27, 29, 31–33, 39–42, 65, 66, 87, 101, 102, 121
- Ling** Linguistic Features. 38–40
- LSTM** Long Short-Term Memory. 35
- MACE** Multi-Annotator Competence Estimation. 53, 54, 60, 98
- MF** Moral Foundation. 3, 4, 8, 14, 16–18, 28–30, 89, 91–93, 95, 96, 99–103, 121
- MFD** Moral Foundations Dictionary. 18, 28–30, 92, 93, 95–97, 100, 101, 121

- MFT** Moral Foundations Theory. 3, 8, 9, 16–18, 29, 89, 100, 102
- MFTC** Moral Foundations Twitter Corpus. 95–97
- MT** Moral Trait. 17, 18, 91, 92, 101
- mturk** Amazon Mechanical Turk. 45, 52, 76
- NLP** Natural Language Processing. 9, 12–14, 20, 48
- NN** Neural Network. 36, 38–41
- NP** Noun Phrase. 25, 27, 67, 122
- O** Object. 45, 46, 49–51, 57, 58
- OpenIE** Open Information Extraction. 26, 27, 73–76, 80, 82, 83
- P** Predicate. 45, 46, 49–51, 57, 58
- PABAK** Prevalence- and Bias-Adjusted Kappa. 95
- POS** Part of Speech. 12, 15, 16, 48, 59
- PPR** Personalized PageRank. 16, 42, 92, 101
- RoBERTa** Robustly Optimized BERT Pretraining Approach. 14, 59–61
- SBERT** Sentence-BERT. 14, 93, 94, 96, 98
- sent** Sentiment. 45, 46, 49–51, 55, 56
- StArCon** Stance Detector for Arguments from Consequences. 21–24, 26, 29, 55–57, 59–63, 66–68, 83, 121
- T** Target. 45–48, 50, 51, 57, 58
- tf-idf** Term Frequency-Inverse Document Frequency. 71
- WN-PPR** WordNet extended using Personalized PageRank. 92, 95–97

Chapter 1

Introduction

Argumentation is a particularly challenging means of communication as usually its goal is to change the audience’s opinion. While the phenomenon of argumentation is researched since centuries, there exists little work on the automatic explanation of arguments. General approaches for explaining texts, such as linking difficult or unknown concepts to their definition or a description, are an important step also towards explaining arguments, but they lack to complete the coherence of arguments, or only accidentally achieve to do so. Instead, we aim to specifically explain arguments in a formalized and well-defined way which facilitates downstream tasks such as (counter-)argument generation or large scale analyses of debates.

Given the complexity of the phenomenon argumentation and its widespread use among different domains such as politics, court, mathematics, medicine or every day communication, this task is very challenging. Consequentially, many works on analyzing or mining arguments focus on a particular domain (i.e., Mochales and Moens, 2011; Stab and Gurevych, 2017a; Dusmanu et al., 2017; Green, 2018; Mayer et al., 2020). In contrast, our main focus is on a specific type of arguments.

In the last decades, different ways of schematizing arguments have emerged. One of these schemes is called the *argument from consequences*. An argument from consequences suggests that one should or should not take a certain action because of its potential consequences. Arguments from consequences have caught our special interest due to their very high frequency in online debates, for instance on Debatepedia¹. An annotation study of ours indicates that roughly half of the arguments there involve consequences (see chapter 4.3), and Al-Khatib et al. (2020) annotated cause-effect-relations that are characteristic for arguments from consequences in 1736 out of 4740 claims. Im-

¹debatepedia.org; Unfortunately, the website is offline at the moment of writing this thesis.

portantly, arguments from consequences are defeasible by nature. This means that one can agree to everything which is stated in the argument, but still disagree with its conclusion. Thus, one can often make arguments from consequences both for and against taking a certain action. For instance, it can be argued that *prohibiting motorized individual transport reduces the fine dust pollution in cities* as well as that *prohibiting motorized individual transport will restrict people's freedom of movement*. Both are arguments from consequences about the same action, and both suggest a different conclusion.

The focus on arguments from consequences allows us to access the underlying reasoning of such arguments and address otherwise difficult tasks with explainable methods. One such task is called *enthymeme reconstruction*: A common phenomenon in argumentation, which we frequently observe in arguments from consequences, is that parts of the argument which are important for its coherence are not stated explicitly. Such arguments are often called enthymemes. The task is to make these missing parts explicit. So far, to the best of our knowledge, there exist no approaches towards automatic enthymeme reconstruction that go beyond choosing from a set of given premises (Boltužić and Šnajder, 2016; Habernal et al., 2018). However, we argue that by focusing on a specific argumentation scheme, automatic enthymeme reconstruction becomes feasible (see also Razuvayevskaya and Teufel, 2017). In arguments from consequences, for instance, it is common to state that a certain action might have certain consequences, but it is left implicit whether the consequences are desirable or not and whether the action should be brought about or not.

Example 1. *Unjustified regular intake of dietary supplements can lead to kidney stones and other side effects.*

In the example above, a consequence is expressed quite explicitly. For humans, the argument is easy to understand. However, for the argument to make sense, we need to know that *kidney stones and other side effects* are undesirable which prompts the conclusion that one should not regularly use dietary supplements without reason. We address this task in chapter 4 by concretely identifying the action, the consequence and the consequence's polarity which then enables us to reconstruct the conclusion. For determining the consequence's polarity, we aim to split it into an effect (*can lead to*) and an object (*kidney stones*). This way, we represent the argument as what we call an *effect triple* like $\langle \text{Dietary supplements, can lead to (+), kidney stones (-)} \rangle$.

To evaluate whether we correctly reconstructed the consequence's polarity and the conclusion, we address the task of *stance detection*. Given a topic, like *dietary supple-*

ments, and an argument like example 1, the task consists in determining whether the argument is in favor or against the topic. Stance detection currently receives a lot of attention (Küçük and Can, 2020; ALDayel and Magdy, 2021). We propose an unsupervised stance detection method which is based on the effect triple. What sets our method apart from related work is that it focuses on one particular scheme of arguments and offers the effect triple as an explanation for its classification result. Unlike many traditional approaches (i.e., Somasundaran and Wiebe, 2010; Anand et al., 2011; Hasan and Ng, 2013; Faulkner, 2014; Sobhani et al., 2016; Addawood et al., 2017; Sun et al., 2018; Du et al., 2017; Dey et al., 2018; Ghosh et al., 2019), our method is topic independent and involves no training. Modern transformer based approaches outperform our method (see chapter 4.5), but they lack the explainability which our method offers.

Building upon the reconstruction of the argument, we aim to explain the two key aspects of the argument: First, we offer explanations for why the action has the postulated consequence. Second, we address why the consequence is considered desirable or not. For the first type of explanation, we exploit the transitivity of *effect relations* such as *dietary supplements* $\xrightarrow{+}$ *kidney stones*. We extract such effect relations from large argumentative and encyclopedic text resources and build a Knowledge Graph (KG) with them which we call *effect graph*. Such a KG can be useful not only for explaining arguments, but also for other tasks such as large scale debate analyses, extending common-sense KGs, or argument retrieval. In that regard, our work is similar to Al-Khatib et al. (2020) and Al Khatib et al. (2021), who also build a KG based on effect relations extracted from text. Other than them, we propose a precision focused unsupervised effect relation extraction method. We improve upon their KG by also including effect relations from encyclopedic texts, by having a substantially larger KG, and by making the effect graph publicly available. Using the effect graph, we explain arguments from consequences by chaining effect relations or combining them with lexical knowledge. For instance, to explain why *dietary supplements* can have negative impacts on *health*, we might use the effect relation outlined above (*dietary supplements* $\xrightarrow{+}$ *kidney stones*) along with a second effect relation stating that *kidney stones* have negative consequences on health.

For explaining why the consequence is considered desirable or not, we make use of the Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Graham et al., 2013). The theory claims that there exist few Moral Foundations (MF) which are the basis for intuitive ethics across different cultures (see chapter 2.1.4). These MFs are *care*, *fairness*, *loyalty*, *authority*, *purity* and, potentially, *freedom*. The human annotation and

automatic classification of MFs is being researched actively (i.e., Johnson and Goldwasser, 2018; Hoover et al., 2020; Hopp et al., 2021). We also train MF classifiers and, in contrast to related work, concretely evaluate their use on arguments. Using the best performing classifier and manual annotations, we explore the usage of MFs in argumentative texts. We find their usage to be weakly correlated with argument quality and audience approval.

Concluding, we propose methods to model and explain several aspects of arguments from consequences in a structured way. Generally, we believe that at least for defeasible arguments, finding different solutions for different types of arguments can not only improve classification results, but also make otherwise unfeasible tasks such as enthymeme reconstruction possible.

In the following, we formally define our research problem in section 1.1. We list our contributions in section 1.2 and outline the structure of this thesis in section 1.3.

1.1 Problem Definition

The primary goal of this thesis is to research the automatic explanation of arguments from consequences. We define them as described in Walton (1999):

Definition 1 (Argument from consequences). *“The argument for accepting the truth (or falsity) of a proposition by citing the consequences of accepting that proposition (or of not accepting it).”*

In logics, this argument is often treated as a fallacy because the consequences of a thesis are irrelevant for its acceptance (Rescher, 1964; Fischer, 1971). Still, Walton (1999) highlights that this type of argument can be quite reasonable, e.g., when arguing for or against a proposed policy, and that it is widely used in everyday argumentation, which fits our own observation.

Indeed, arguments from consequences are intuitively easier to accept if they are formulated similarly as proposed in Walton et al. (2008) who formalizes four different subtypes of arguments from consequences. The first two can be collapsed as follows:

Definition 2 (Argument from positive (*negative*) consequences).

Premise *If A is brought about, good (bad) consequences will plausibly occur.*
Conclusion *Therefore, A should (not) be brought about.*

While the argumentation scheme is rationally compelling when presented this way, it is still defeasible. Rationally, it is well possible to accept the argument's premise but not its conclusion. This, however, is neither uncommon nor does it make the scheme obsolete. Presumably, most debates are conducted with defeasible arguments since the existence of a non-defeasible, i.e., deductively valid argument would solve the debate.

Oftentimes, in argumentation, premises or conclusions that are obvious are left implicit. Therefore, a typical argument from consequences may be formulated as follows:

Example 2. *Investing into renewable energies reduces the CO2 emissions.*

Not only is the argument's conclusion, that one should invest in renewable energies, implicit, but it is also not explicitly stated whether the described consequences are good or bad. While Walton's formalization is perfectly reasonable from an argumentation theoretical point of view, we propose a different notation in order to facilitate operationalization and analysis²:

Notation 1 (Argument from consequences).

Effect premise	<i>If ACTION is brought about, CONSEQUENCE will plausibly occur</i>
Judgment premise	<i>CONSEQUENCE is good (bad)</i>
Conclusion	<i>ACTION should (not) be brought about</i>

In terms of this notation, example 2 merely consists of the effect premise while the judgment premise and the conclusion are left implicit. While this is the most common case which consequently we will put our main focus on, one can also think of arguments where the effect premise is implicit:

Example 3. *We should invest into renewable energies because CO2 emissions threaten to destroy our planet.*

This example consists of the conclusion (*We should invest into renewable energies*) and the judgment premise which suggests that CO2 emissions are bad. The latter, however, is not formulated explicitly, but it is the conclusion of yet another argument from

²The proposed notation is in line with several web sources, but we could not trace it back to its origin. Some of these web sources are:

- https://en.wikipedia.org/wiki/Appeal_to_consequences
- <https://yandoo.wordpress.com/2014/01/26/argument-from-consequences/>
- <https://fallacyinlogic.com/appeal-to-consequences/>
- https://loricism.fandom.com/wiki/Argument_from_Consequences
- https://rationalwiki.org/wiki/Appeal_to_consequences

Reconstruction of example 3		
<i>Effect premise</i>	Investing into renewable energies negatively affects CO2 emissions.	<i>implicit</i>
<i>Judgment premise *</i>	CO2 emissions threaten to destroy our planet.	<i>explicit</i>
<i>Conclusion</i>	We should invest into renewable energies.	<i>explicit</i>
Reconstruction of the judgment premise *		
<i>Effect premise</i>	CO2 emissions threaten to destroy our planet.	<i>explicit</i>
<i>Judgment premise</i>	Destroying our planet is bad.	<i>implicit</i>
<i>Conclusion</i>	We should not produce CO2 emissions.	<i>implicit</i>

Table 1.1: Reconstruction of example 3 and its inner argument.

consequences consisting only of the effect premise (*CO2 emissions threaten to destroy our planet*). The reconstructions of both example 3 and its inner argument are shown in table 1.1. As a side note, this example also demonstrates why the terms conclusion and premise are relative.

Summarized, the first problem we face when dealing with arguments from consequences is implicitness of the premises or the conclusion:

Problem 1. *The conclusion and either the effect premise or the judgment premise can be implicit.*

Our next problem is a computational linguistic one:

Problem 2. ACTION and CONSEQUENCE are to be identified.

This problem is quite intuitive: As arguments from consequences are usually stated in natural language and not in our or Walton's formalized form, we need to identify the two core instances ACTION and CONSEQUENCE in order to be able to properly analyze and explain the argument.

Further, we aim to clarify the effect premise and the judgment premise. Concerning the effect premise, assuming ACTION and CONSEQUENCE are identified, we formulate the following problem:

Problem 3. *In the effect premise, it might be unclear why ACTION might cause CONSEQUENCE.*

“Why” refers to the rationale rather than the motive. In terms of our examples, this means to explain why or how investing in renewable energies reduces the CO2 emissions or, respectively, why CO2 emissions threaten to destroy our planet.

Similarly, the judgment premise might lack further explanation:

Problem 4. *Concerning the judgment premise, it might be unclear why CONSEQUENCE is considered good or bad.*

Considering example 2, the question would be why reducing CO2 emissions is considered good. For example 3, the question why CO2 emissions are bad is already answered explicitly (because they threaten to destroy our planet). However, we can further ask why destroying our planet is bad to properly explain the inner argument.

Note that if an argument from consequences is formulated explicitly, including effect premise, judgment premise and conclusion, then we do not see any further need to explain why the conclusion follows from the premises as this is exactly the concept of the argumentation scheme.

Summarized, we address the following task: When presented an argument from consequences, e.g., from an online debate, we reconstruct the argumentation scheme which involves identifying ACTION, CONSEQUENCE and CONSEQUENCE's alignment. Further, we offer explanations for why ACTION might cause CONSEQUENCE and for why CONSEQUENCE is considered good respectively bad. The reconstruction of the scheme as well as the proposed explanations are in a structured format which allows for both further processing as well as generating an answer in natural language.

1.2 Contributions

Our main research contributions consist in creating methods and data for analyzing and explaining arguments. In the following, we list these contributions categorized by the problem they address.

Problem 1 For reconstructing the judgment premise and the conclusion,

- (i) we propose the concept of effect triples, which generalizes and extends the templates from Reisert et al. (2018), in order to model the most relevant aspects of an argument from consequences;
- (ii) we create a method to automatically annotate effect premises with effect triples, which effectively reconstructs the judgment premise and the conclusion;
- (iii) for evaluating our argument reconstruction, and as a contribution to the field of stance detection, we propose an intuitive method to detect the stance of arguments from consequences based on the effect triples;
- (iv) we create and share a corpus of arguments with crowd-annotations about stance and whether the arguments refer to consequences or not (chapter 4).

Problem 2 The effect triples mentioned above also address the problem of concretely identifying ACTION and CONSEQUENCE. Further, we propose a method to extract effect relations which is more focused on precision than related work (chapter 5).

Problem 3 For explaining why ACTION might lead to CONSEQUENCE,

- (i) we build a KG of effect relations which, in contrast to related work, is larger, contains effect relations from argumentative and encyclopedic resources, and is publicly available;
- (ii) we extensively evaluate this KG, for which we create and share annotations to access its precision and recall;
- (iii) we are the first to propose a method to generate and rank explanations for effect relations (chapter 5).

Problem 4 In order to explain why CONSEQUENCE, or any concept, is considered good or bad, we suggest to apply the MFT. Our contributions include

- (i) the first annotation of MFs on a dataset of arguments;
- (ii) the evaluation of different supervised models for the classification of MFs in tweets and, newly, specifically in arguments;
- (iii) an exploratory study about the use of MFs in online debates (chapter 6).

1.3 Outline

The remainder of this thesis is structured as follows: In chapter 2 we present the required theoretical background and position our research within related work. In chapter 3 we present preliminary work in which we classify argumentative relations with background knowledge. In chapter 4 we model and reconstruct arguments from consequences and propose a method for detecting their stances. Afterwards, we separately address the explanations of the two premises of an argument from consequences: In chapter 5 we propose a method for explaining the effect premise, while in chapter 6 we address the classification of MFs and their use in argumentation. We conclude the thesis in chapter 7.

Chapter 2

Theoretical Background and Related Work

2.1 Theoretical Background

In this section, we introduce basic concepts and set the terminology which we will use frequently in the course of the thesis. We specifically address the areas of Argumentation (section 2.1.1), Natural Language Processing (section 2.1.2), Knowledge Graphs (section 2.1.3), and the Moral Foundations Theory (section 2.1.4). We further require the reader to have a basic understanding of Machine Learning and Statistics.

2.1.1 Argumentation

Definition 3 (Argumentation). *“Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint.”*

This definition is from Van Eemeren and Grootendorst (2004). It is also followed by Stede and Schneider (2018) as it compactly unifies the most important aspects of argumentation. Importantly, the term *argumentation* denotes an activity. For the most part, what we in fact analyze is the this activity’s result and not the activity itself. For this purpose, we further define:

Definition 4 (Argument). *An argument consists of a set of premises which supports a conclusion.*¹

¹Note that especially in the context of computational argumentation, it is sometimes stated that

In some cases, as in Aristotle's famous syllogism (see example 4), the conclusion does logically follow from the premises, such that it would be inconsistent to accept the premises but not the conclusion:

Example 4.

Premise 1 All men are mortal.
Premise 2 Socrates is a man.
Conclusion Therefore, Socrates is mortal.

In contrast, there exist arguments where the premises do not imply the conclusion, but only prompt it. We call these arguments *defeasible*, as there might exist sets of premises which support, or possibly imply, an opposing conclusion. A famous example are inductive arguments:

Example 5. *I have seen hundreds of swans, and all of them were white. Thus, all swans are white.*

Even though inductive arguments clearly are defeasible, that does not make them unreasonable per se. Presumably, most debates consist primarily of defeasible arguments as otherwise, there would be no rational reason to continue the debate. We define a debate as follows:

Definition 5 (Debate). *A debate is a set of arguments relating to the same topic.*

The topic can be rather wide, like *death penalty*, or very concrete, like *Introduction of the death penalty for child rapists in Virginia*. Arguments within a debate can be either *in favor* of the topic, *against* the topic or neither of both. We call these polarities the argument's *stance*. For most parts in this thesis, we only consider positive (in favor) and negative (against) stances.

Lastly, we want to highlight that the terms *premise* and *conclusion* are used relatively. It is very well possible that a statement which is the conclusion of an argument is at the same time a premise for another argument. In that regard, we define the following two relations between arguments:

Relations between Arguments

Definition 6 (Support). *Argument A supports argument B iff A's conclusion is a premise for B's conclusion.*

premises either support or attack the conclusion. This, however, contradicts the original idea of premises and we decided not follow this alternative definition. Instead, we see a premise which *attacks* a conclusion as a premise which in fact supports a different, opposing conclusion which might be left implicit.

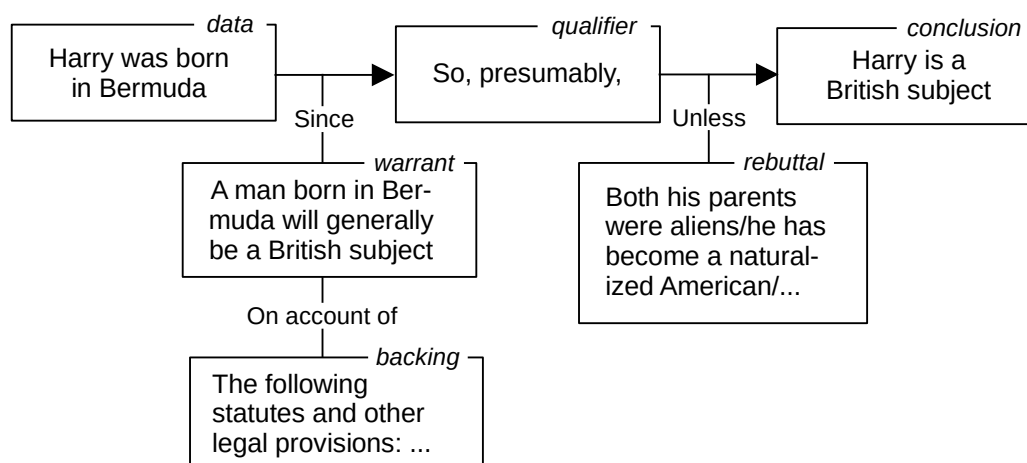


Figure 2.1: Example of the Toulmin model of argumentation. Source: Toulmin (2003).

Definition 7 (Attack). *Argument A attacks argument B iff A's conclusion is in contradiction with argument B.*

The notion of arguments attacking each other is a relevant concept not only for invalidating or weakening, but also for validating arguments: Dung (1995) suggests a framework in which the logical acceptability of an argument depends on whether it can be defended against attacking arguments. Concretely, he calls an argument *A* acceptable with respect to a set of arguments *S* iff for all arguments *B* attacking *A* there exists an argument in *S* which in turn attacks *B*. Although we do not directly work with Dung's framework, it is a nice further motivation for building the effect graph which can also be used to identify attacking arguments.

Models of Argumentation

Although we chose relatively simple definitions to work with, there exist more complex models of how arguments are structured and how they can be explained. One of the most famous models of arguments is from Toulmin (2003). It suggests that arguments consist of six components, one of which is the *conclusion* which is mainly supported by the *data*. The *warrant* provides reason for how the data supports the conclusion, and the warrant itself might be further supported by some *backing*. Arguments further contain a *qualifier* which indicates the strength of the warrant and, importantly, a *rebuttal* which presents circumstances in which the warrant is not valid. Figure 2.1 shows an example where the model is applied.

While the Toulmin model itself is general and applicable to all arguments, it inspired the emergence of very concrete argumentation schemes which essentially group

and identify arguments by their warrant. The possibly most complete collection of argumentation schemes is presented in Walton et al. (2008). They define argumentation schemes as follows:

Definition 8 (Argumentation Scheme). “*Argumentation Schemes are forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation.*”

There exist schemes for both logically valid and defeasible arguments. In their user’s compendium of schemes, Walton et al. (2008) present sixty primary schemes, potentially having subschemes. Each of these schemes consists of a set of premises and a conclusion, containing placeholders as shown in definition 2. Further, they come along with a set of critical questions. For the argument from consequences, these questions are:

CQ1: “How strong is the likelihood that the cited consequences will (may, must) occur?”

CQ2: “What evidence supports the claim that the cited consequences will (may, must) occur, and is it sufficient to support the strength of the claim adequately?”

CQ3: “Are there other opposite consequences (bad as opposed to good, for example) that should be taken into account?”

Note that the critical questions resemble components of the Toulmin model: CQ1 is about *qualifiers*, CQ2 about the warrants *backing*, and CQ3 at least somewhat represents a *rebuttal*.

2.1.2 Natural Language Processing

Methodologically, the most relevant research area for this thesis is Natural Language Processing (NLP). In the following, we introduce relevant tasks and concepts.

Part of Speech Tagging Part of Speech (POS) Tagging is the task to assign each word in a sentence or phrase its according POS (noun, verb, article, ...). It is fundamental for many subsequent tasks, but we also use it explicitly at some times.

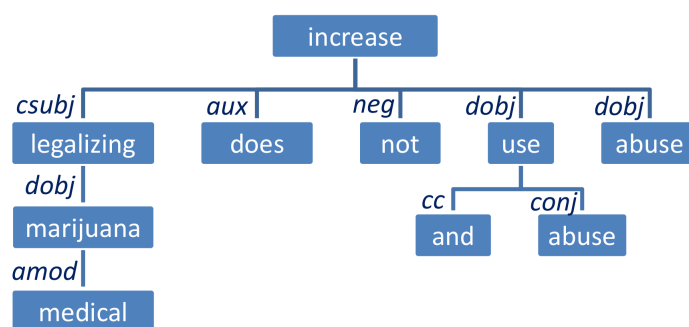


Figure 2.2: Example dependency graph.

Lemmatizing and Stemming Lemmatizing deals with a word’s inflectional variance by transforming the word into its lemma. For example, *loves* and *loving* both become *love*, while *lovers* becomes *lover*. Stemming on the other hand transforms a word into its stem. For example, *loves*, *loving* and *lover* would all be turned into *lov*. However, when compared to lemmas, stems have the risk of being ambiguous: For instance, the stems of the words *caring* and *cars* both are *car*, while their lemmas are *care* and *car*.

Dependency Parsing In dependency parsing, a sentence is transformed into a directed graph (oftentimes a tree) where each node corresponds to a word in the sentence and the edge indicates a specific dependency relation. Figure 2.2 shows a dependency graph for the sentence *Legalizing medical marijuana does not increase use and abuse*. We mainly use dependency parsing to determine the subject and object of an effect verb such as *increase* and to detect negations. The dependency relations which are relevant for this thesis are described in appendix B.

We use the Stanford CoreNLP Natural Language Processing Toolkit (Manning et al., 2014) for dependency parsing as well as the previously described NLP tasks.

Sentiment Analysis Medhat et al. (2014) define sentiment analysis as follows:

Definition 9 (Sentiment Analysis). *Sentiment Analysis [...] is the computational study of people’s opinions, attitudes and emotions toward an entity.*

Usually, sentiment analysis is considered a classification task where the output is a *sentiment score*, often between -1 and 1, or a *sentiment polarity* (positive, neutral, negative). We often refer to instances expressing a positive sentiment, such as *love*, as being *good* while we refer to instances expressing a negative sentiment as being *bad*. Medhat et al. (2014) distinguish between three different levels of sentiment analysis: document,

sentence, aspect. We explicitly use sentence- and aspect-level sentiment analysis tools and methods in various contexts in this thesis. We use a sentence’s sentiment as a feature in chapter 3. In chapter 4, we again use sentence-level sentiment analysis as a baseline for our stance detection method. Further, in order to classify CONSEQUENCE’s polarity, we use *sentiment lexicons* which assign sentiment scores to words.

Language Models In modern NLP, it is common to transform textual input into a vector space. This process as well as the process’ result is then called *embedding*. While there exist quite simple approaches such as Word2Vec (Mikolov et al., 2013a,b) or GloVe (Pennington et al., 2014) to embed single words, it recently got popular to use transformer encoders which also consider the context a word appears in when embedding it. The most prominent language model used in this thesis is called *Bidirectional Encoder Representation from Transformers (BERT)* (Devlin et al., 2018). Essentially, BERT was pretrained for masked language modeling and next sentence prediction on a large resource of texts. It is expected to entail a solid understanding of linguistic patterns through this pretraining and it can further easily be fine-tuned to serve as a classifier for a large variety of NLP tasks. Fine-tuning means that one additional softmax layer is added to the end of BERT’s pipeline and the entire network is trained again for some epochs. Since only the newly added layer has to be trained from scratch, fine-tuning is relatively efficient.

Nowadays, fine-tuned pretrained transformers receive the best classification results in most NLP tasks (Lin et al., 2022). Besides BERT, we compare our stance classifier to its two successors *Robustly Optimized BERT Pretraining Approach (RoBERTa)* (Liu et al., 2019) and *Decoding-enhanced BERT with Disentangled Attention (DeBERTa)* (He et al., 2021) in chapter 4.5. Further, in chapter 6, we experiment with using *Sentence-BERT (SBERT)* (Reimers and Gurevych, 2019) for MF classification. Generally, SBERT encodes sentence similarity in a human interpretable way such that similar sentence pairs can be retrieved efficiently, based on cosine similarity. However, given an anchor text, a positive, and a negative text sample, SBERT can also learn to maximize a score based on the anchor’s similarity to the positive sample and its distance to the negative sample.

2.1.3 Knowledge Graphs

In this thesis, we use existing KGs as resources of external knowledge, and we build our own KG, the effect graph, to structure external knowledge from textual resources. We

define KGs as proposed in Hogan et al. (2021):

Definition 10 (Knowledge Graph). *A Knowledge Graph (KG) is “a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities”.*

In the following, we briefly introduce the already existing KGs we use. All these KGs are available in multiple languages, but we exclusively use the English versions.

DBpedia DBpedia (Auer et al., 2007; Lehmann et al., 2015) is a comparably large KG with millions of entities. It contains structured information from Wikipedia such as hyperlinks between Wikipedia articles.

WordNet WordNet (Miller, 1995; Fellbaum, 2010) specifically contains lexical knowledge. The entities in WordNet are so called *synsets*, which contain one or several words or short phrases being synonyms of each other. However, from a word’s perspective, each synset the word is part of represents one of the word’s different meanings. This way, in WordNet, all words are disambiguated. The edges in WordNet are semantic relations, like *hypernym*, *meronym*, *entailment* or *antonym*.

ConceptNet ConceptNet (Speer et al., 2017) is a KG capturing common sense knowledge. Its entities are called *concepts* which consist of potentially POS-tagged words or short phrases. We use the English version of ConceptNet 5.6 which consists of 1.9 million concepts. The edges indicate relatedness between the two concepts. They are labeled with one of 37 relations, some of which are commonly used in other resources like WordNet (e.g., *IsA*, *PartOf*) while most others are more specific to capturing commonsense information and as such are particular to ConceptNet (e.g., *HasPrerequisite* or *MotivatedByGoal*). The most common relation, unfortunately, is the unspecific relation *RelatedTo*.

As we are working with textual data for most of the time, one key challenge is to identify the relevant entities in the KGs. While we use pragmatic solutions most of the time, we want to briefly introduce the corresponding research task (Shen et al., 2015):

Definition 11 (Entity Linking). *“Entity Linking is the task to link entity mentions in text with their corresponding entities in a knowledge base.”*

Obviously, the difficulty of entity linking greatly depends on the KG to link the entity mentions to. For instance, when linking to ConceptNet, one basically needs to perform POS-tagging and string matching, while linking to WordNet or DBpedia on the other hand requires proper word sense disambiguation (Navigli, 2009; Bunescu and Pasca, 2006; Bevilacqua and Navigli, 2020), for instance to distinguish the city *Paris* from the mythological character.

Lastly, we introduce an important algorithm not only for KGs, but for all kinds of networks: The node centrality measure PageRank (Page et al., 1999). In general, node centrality measures are means to quantify how important certain nodes are. There exist quite simple measures, e.g., based on a nodes degree or its distance to the other nodes in the graph. PageRank was originally designed to rank websites based on their importance. From a graph theoretic perspective, PageRank’s main idea is that nodes are more important the more incoming edges from important neighbors they have. In addition to the traditional PageRank measure, there exist variations of it, one being especially relevant for us: Personalized PageRank (PPR) (Haveliwala, 2002). In this variant, one specifies a set of seed nodes which consecutively get free importance, thus making their outgoing neighbors more important, and so forth. We refer to it in chapter 3 and in chapter 6 where we use it to identify nodes in WordNet that are relevant with respect to specific Moral Foundations.

2.1.4 Moral Foundations Theory

In order to explain the judgment premise (problem 4), i.e., to explain why something is considered good or bad, we need to consider human values or morals. To operationalize the concept of morality, we use the Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Graham et al., 2013). According to the theory, there exist “several innate and universally available psychological systems [which] are the foundations of ‘intuitive ethics’”². In detail, the MFT consists of four claims which we will briefly describe in the following:

- **Nativism:** The theory claims that there exists a “‘first draft’ of the moral mind” (Graham et al., 2013) when we are born.
- **Cultural learning:** This first draft gets further revised through experience, which is the reason for cultural differences concerning morality.

²<https://moralfoundations.org>

- **Intuitionism:** This claim is based on the Social Intuitionist Model (Haidt, 2001) which states that “moral judgment is generally the result of quick, automatic evaluations (intuitions)”.
- **Pluralism:** As throughout human evolution there were many recurring social challenges, consequently there are multiple MFs.

In the following, we briefly introduce the proposed MFs and their origin. Each MF is referred to by a *virtue* as well as an opposing *vice*.

- **Care/Harm:** Caused by the need to protect and care for children, this foundation is about the virtues of caring and kindness and triggers emotions such as compassion for victims and anger at perpetrators.
- **Fairness/Cheating:** Individuals which are sensitive to the ideas of cheating and cooperation had an advantage when presented with opportunities to engage “non-zero-sum exchanges and relationships”. This foundation is about virtues such as fairness, justice and trustworthiness.
- **Loyalty/Betrayal:** Humans needed to form cohesive coalitions to survive, and today this foundation is still triggered by nations, sports teams etc. Relevant virtues are loyalty, patriotism and self-sacrifice.
- **Authority/Subversion:** Basically, this foundation is about hierarchies. Relevant virtues are obedience and deference.
- **Sanctity/Degradation:** This foundation has the very practical origin of avoiding communicable diseases and is tied to the emotion of disgust. It is also sometimes referred to as *Purity/Degradation*. The relevant virtues involve temperance, chastity, piety and cleanliness.

Graham et al. (2013) emphasizes that they expect disagreement about this particular list of MFs. In fact, Haidt and Joseph (2004), who initially introduce these foundations (under slightly different names) consider *Liberty/Oppression* to be another candidate for a MF (Iyer et al., 2012).

Sometimes it is helpful to consider the virtue and the vice of a MF to be different categories. For this purpose, we refer to them as Moral Trait (MT). Thus, *care* and *harm* would be considered different MTs while belonging to the same MF.

Though we will stick to the MFT throughout this thesis, we want to briefly mention another interesting concept in this context: *human values*. There exist several different definitions. A particularly simple summation of them is from Cheng and Fleischmann (2010):

Definition 12 (Human Values, Cheng and Fleischmann). *Human values “serve as guiding principles of what people consider important in life”.*

One important theory about human values which we initially considered as an alternative to the MFT is the Schwartz Value Theory (Schwartz and Boehnke, 2004). Schwartz (1994) defines human values based on five features commonly agreed on in literature:

Definition 13 (Human Values, Schwartz). *“A value is a (1) belief (2) pertaining to desirable end states or modes of conduct that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities.”*

The Schwartz Value Theory suggests that there are ten basic values which are common to all societies. These values are: *power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity, security*. Even though these values differ from the MFs, the Schwartz Value Theory resembles the MFT, and indeed the relation between the two theories is being researched (Zapko-Willmes et al., 2021). We chose to work with the MFT mainly because of the distinction of virtues and vices and the existing Moral Foundations Dictionary (MFD) which contains a list of relevant words for each MT.

2.2 Related Work

In this section, we first discuss related work for explaining arguments in general in section 2.2.1. In section 2.2.2, we specifically address stance detection and related tasks, which is relevant especially for chapters 3 and 4. In section 2.2.3, we discuss relation extraction and KG generation, which is relevant for chapters 4 and 5. Lastly, we introduce related work for MF classification in section 2.2.4 which relates to chapter 6. We conclude by summarizing our research contributions in section 2.2.5.

2.2.1 Explaining Arguments

While there exists much research about explaining a model’s predictions, *Explainable Artificial Intelligence* being a research field on its own (Adadi and Berrada, 2018), there is comparably little research about explaining texts such as, in our case, arguments. Part of the reason might be that *explaining* is a broad term. There exist many different

problems a reader might have with a text, so consequentially, there exist many different explanations. One potential problem a reader might have with a text is that the text is linguistically too complex. In this case, one could consider *text simplification* (Siddharthan, 2014; Al-Thanyyan and Azmi, 2021) to be an explanation. Further, the text might contain words or entities which are unknown to the reader. Then, *entity linking* (Shen et al., 2015; Özge Sevgili et al., 2022) to a lexicon or encyclopedia might provide an adequate explanation. Though extremely important and useful, these are not the kind of explanations we are interested in. Instead, we aim to explain arguments in a way which is exclusive to arguments: By making them coherent.

Obviously, most arguments we face are meant to already be coherent. But, more often than not, these arguments do not explicitly state everything which is needed to draw the conclusion, but leave it to the reader to fill in the gaps. Our task is then to make these missing bits explicit, which is why we introduced the notion of *argument explicitation* (Hulpuş et al., 2019; Becker et al., 2020). In the following, we discuss two steps which are important in that regard: *argumentation scheme identification* and *enthymeme reconstruction*.

Argumentation Scheme Identification

The task of argumentation scheme identification consists of two parts: classifying the argumentation scheme and identifying the scheme’s components in the argument. Concerning Walton Schemes, there exists some research about the former task (Walton and Macagno, 2015; Palau and Moens, 2009; Feng and Hirst, 2011).

However, in the context of this thesis, we are particularly interested in the task of identifying a scheme’s components in the argument. In our case, this means to identify not only the effect premise, the judgment premise and the conclusion, but also to concretely identify ACTION, CONSEQUENCE, as well as CONSEQUENCE’s and the conclusion’s polarities (see notation 1). Reisert et al. (2018) refer to this task as *argument template instantiation*. We address this task in chapter 4 where we use a lexicon-based unsupervised method to extract effect triples from arguments from consequences. Similarly to us, Reisert et al. (2018) focus mainly on arguments from consequences. They propose eight templates to describe arguments from consequences. They describe CONSEQUENCE as a positive or negative effect over an object which is good or bad. Each of their templates addresses one combination of the polarities of the effect, the object, and the conclusion (whether ACTION should be brought about or not). Half of these templates are described as *support templates* and the other half as *attack templates*. Our

effect triples can be seen to be a generalization and extension of these templates. Besides proposing the templates, Reisert et al. (2018) perform an annotation study with promising Inter-Annotator Agreement (IAA). Unlike Reisert et al. (2018), we do not only propose the effect triples to instantiate arguments from consequences, but also address the task computationally.

Enthymeme Reconstruction

Definition 14 (Enthymeme). *An enthymeme is an argument where at least one premise or the conclusion is not stated explicitly.*³

The task of enthymeme reconstruction consists in reconstructing the argument by making the missing premises and conclusion explicit, which directly relates to our problem 1, but also to problems 3 and 4. In the context of argument mining, this task is also sometimes called *argument completion* (Peldszus and Stede, 2013). Note however that at least for arguments from consequences, making the conclusion explicit is very much related to solving the stance detection task (see section 2.2.2): Identifying the stance of the argument towards ACTION makes it trivial to conclude whether ACTION should be brought about or not.

While there exists some theoretical work on enthymeme reconstruction (Walton and Reed, 2005; Paglieri and Woods, 2011; Black and Hunter, 2012; Hosseini et al., 2014), there is only little applied work in the NLP domain. Rajendran et al. (2016) make a step towards enthymeme detection by classifying stances to be either explicitly or implicitly expressed. Similarly, Stab and Gurevych (2017b) automatically identify insufficiently supported arguments. However, they are not specifically identifying enthymemes, but arguments which do not fulfill the argument quality criterion of *sufficiency* (see also Habernal and Gurevych, 2016; Wachsmuth et al., 2017b).

As a step towards automatically reconstructing enthymemes, Boltužić and Šnajder (2016) have humans annotate missing premises which bridge the gap between a premise and a conclusion. They find that the number of missing premises correlates negatively with the textual similarity between the provided premise and conclusion. Consequently, in their experiments for automatically reconstructing enthymemes, they choose one or more premises from a set of premises such that, when included, the said textual similarity increases. Further, Habernal et al. (2018) organized a shared task which they

³Note that this definition of enthymemes is inconsistent with the original notion from Aristotle as pointed out by Burnyeat (1994). Walton and Reed (2005) instead propose the term *incomplete argument*, but they also note that “tradition, especially one so well-entrenched as this one, is hard to change”.

call *Argument Reasoning Comprehension Task*. Essentially, one is provided a premise, a conclusion and two possible warrants and the task consists in selecting the warrant which makes the argument coherent. Both approaches have in common that the missing premises are selected and not generated.

Oftentimes, enthymeme reconstruction is postulated to be subjective (Hitchcock and Hitchcock, 2017; Scriven, 1976; Gough and Tindale, 1985; Burke, 1985) which Walton and Reed (2005) discuss under the term *Attribution Problem*. We avoid this problem by making clear that our aim is *not* to recover what an author meant and did not write down, but to complement or explain the argument at hand in a way which makes sense. Razuvayevskaya and Teufel (2017) show that enthymeme reconstruction can be objectively possible under some circumstances. They annotate enthymemes in what they call *let alone arguments*⁴ with high IAA. While such arguments are not particularly relevant for us, we also do not try to find a one-fits-all solution for the immensely complex task of enthymeme reconstruction, but instead narrow the scope by focusing on one particular type of arguments.

2.2.2 Stance Detection and related tasks

Stance Detection is the task to predict whether a statement is in favor of or against a given topic. In chapter 4, we propose an unsupervised method to detect the stance of arguments from consequences specifically which we denote by *Stance Detector for Arguments from Consequences (StArCon)*. In chapter 3, we extract features from KGs with the goal to improve stance detection and argumentative relation detection.

Stance detection is crucial for us for several reasons: First, it is related to reconstructing the conclusion of an argument and, generally speaking, a key part of explaining an argument. Second, in contrast to most other tasks we address, stance detection is easily quantifiable which is why we use it to evaluate our scheme reconstruction. Third, the task is very useful on its own. Thus, unsurprisingly, there exists much research about stance detection and similar tasks.

⁴Example: *He cannot draw rainbows, let alone unicorns.*

Stance Detection

⁵ Stance detection has been studied on various types of formal texts such as congressional debates (Thomas et al., 2006) and company-internal discussions (Murakami and Raymond, 2010). However, like most recent related work on the topic, we are particularly interested in informal texts from online social media.

Unsurprisingly, BERT has been used successfully for detecting stances in different datasets (Schiller et al., 2020), and it also outperforms our own models in terms of F1 scores. The vast majority of other previous approaches proposes supervised methods, using traditional machine learning algorithms (Somasundaran and Wiebe, 2010; Anand et al., 2011; Hasan and Ng, 2013; Faulkner, 2014; Sobhani et al., 2016; Addawood et al., 2017) and more recently, various deep neural networks architectures (Sun et al., 2018; Du et al., 2017; Dey et al., 2018; Ghosh et al., 2019). These approaches, most of which have been triggered by a recent SemEval shared task⁶ (Mohammad et al., 2016), learn topic-specific models. Thus, new topics require new models whose training entails large user annotation studies. In contrast, both of our proposed methods are topic-independent. In our supervised approach, we try to compensate this by adding features based on background knowledge. StArCon on the other hand is fully unsupervised and rather targets a particular but frequent class of claims, those that refer to consequences.

Among the unsupervised approaches, the most prominent one is of Somasundaran and Wiebe (2009), which got extended by Konjengbam et al. (2018) and Ghosh et al. (2018). However, they focus on non-ideological topics (usually products, e.g., *iPhone* vs. *Galaxy*). In contrast, we target ideological topics (e.g., *Gay Marriage*, *Abortion*) whose stance is harder to detect due to less frequent use of sentiment words and a wider variety of brought up issues and arguments (Rajendran et al., 2016; Wang et al., 2019). On the one hand, Somasundaran and Wiebe (2009), Konjengbam et al. (2018) and Ghosh et al. (2018) extract topic aspects (e.g., *screen resolution*, *battery*) and polarities towards these aspects, a step that is unfeasible for ideological topics. On the other hand, like these works, in StArCon we also use syntactic rules, but not for pairing aspects to opinions, but for extracting triples that correspond to statements about effects over opinion words.

Another class of stance detection approaches uses the context of the post, such as

⁵This subsection is adapted from Kobbe, J., Hulpus, I., and Stuckenschmidt, H. (2020a). Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.

⁶<http://alt.qcri.org/semeval2016/task6>

its relations to other posts in the debate, the network of authors, or the author’s identity (Hasan and Ng, 2013; Sridhar et al., 2014; Addawood et al., 2017; Bar-Haim et al., 2017b). By contrast, we target claim-topic pairs in isolation.

Another aspect that sets our work apart from most related work is that, except for the approaches that target tweets, most focus on longer texts while we consider short, one-sentence claims. In this regard, but not only, the stance detection work that is closest to StArCon is the partly supervised system of Bar-Haim et al. (2017a). They also propose a topic-independent solution to stance detection for short claims without considering context, but they do not specifically address arguments from consequences. While they follow a similar sequence of steps as we do in StArCon, they propose different approaches for each step. For instance, they propose a supervised approach to detect the target of a claim’s opinion, while we do it in an unsupervised manner. They focus primarily on detecting contrastive relations between phrases, while our focus is on detecting effects. In this last regard, the works can be considered complementary.

Sentiment Analysis and Opinion Mining

Both sentiment analysis and opinion mining are tasks which are closely related to stance detection. Unfortunately, they are not defined consistently and are often used interchangeably (Pang et al., 2008; Cambria et al., 2013; Ravi and Ravi, 2015). We define them as follows:

Definition 15 (Sentiment Analysis). *Sentiment analysis is the task to predict whether a document or statement has a positive, negative or neutral sentiment.*

The main difference to stance detection is that in sentiment analysis, there is no provided topic towards which the polarity is expressed. The sentiment is expected to encompass or summarize the polarity for the whole text.

Definition 16 (Opinion Mining). *Opinion mining is the task to extract a target towards which an opinion is expressed, as well as the opinion’s polarity.*

In opinion mining, there also is not provided a topic, but the task involves the extraction of the target towards which the opinion is expressed. In literature, this is often referred to as *aspect-based* opinion mining or sentiment analysis, and more recently as *Target-Stance Extraction* (Li et al., 2023).

Example 6. *It sucks that renewable energies are so underrepresented because fossil fuels are the worst.*

- **Sentiment Analysis:** *negative*
- **Opinion Mining:** *positive wrt renewable energies, negative wrt fossil fuels*
- **Stance Detection wrt renewable energies:** *positive*
- **Stance Detection wrt coal power:** *negative*

Example 6 illustrates the difference between the tasks of sentiment analysis, opinion mining and stance detection. On the first glance, opinion mining seems to be the most challenging one because it also extracts the polarity’s target. But usually, it is a requirement that the target is mentioned explicitly in the text whereas in stance detection, it is common to ask for stances towards unmentioned entities (such as the stance towards coal power in example 6). Sentiment analysis on the other hand is, as we defined it, truly unspecific and not very relevant for our analysis per se. But the research field provides various sentiment lexicons (Khoo and Johnkhan, 2018), which we use to reconstruct the judgment premise, and methods to expand them (Kanayama and Nasukawa, 2006; Qiu et al., 2009), which we imitate to expand a lexicon containing the effects which transitive verbs express on their objects. Concretely, similarly to Badaro et al. (2018), we expand the lexicon by exploiting synsets in WordNet. We slightly adopt their method to achieve a higher precision at the cost of recall.

Opinion mining as we defined it is very much relevant for our stance detection approach. First, because if the topic is mentioned explicitly in the statement whose stance we want to predict, opinion mining might provide a solution. But also, because in StArCon we first identify a target in the statement which relates to the topic and subsequently try to find the stance which is expressed towards that target. Basically, this is an opinion mining task with the slight difference that it is not really a *mining* task, but a classification one.

More and Ghotkar (2016) provide a short overview of different approaches to opinion mining and identify two main directions: supervised learning and lexicon based. The main advantage of lexicon based approaches is that they are less domain specific, which we consider to be important in our envisioned use case. Examples for such approaches are those from Thet et al. (2010), Chinsha and Joseph (2015), Federici and Dragoni (2016): They represent text as dependency trees and use a set of rules to exploit the grammatical dependencies and the prior sentiment scores derived from lexicons in order to determine what Thet et al. (2010) call *contextual sentiment score*. While this is very much what we also do in StArCon, the rules we propose to aggregate sentiment scores differ from and sometimes contradict those of Thet et al. (2010), Chinsha and Joseph (2015), Federici and Dragoni (2016). While our aggregation is conceptually

simpler, we expand upon these works by not only considering sentiment scores, but also effect scores.

This idea, however, is not particularly new: Nasukawa and Yi (2003) concentrate on the semantic relationships between the sentiment expressions and the subject. In addition to nouns, adjectives (which determine the sentiment of a Noun Phrase (NP)) and adverbs (denoting a sentiment towards a verb), they specifically exploit verbs. They distinguish between sentiment verbs that direct a sentiment toward their arguments, and sentiment transfer verbs that transmit sentiments among their arguments. For example, the verb *admire* is positive towards its object while the verb *provide* passes, under certain conditions, the (un-)favorability of its object into its subject (e.g., *XXX provides a good working environment*). They create a sentiment analysis dictionary with such annotations. In chapters 4 and 5, we rely on a similar lexicon: The *connotation frames* (Rashkin et al., 2016) contain, among others, annotations for whether a verb positively or negatively effects its subject or object. We consider that the object’s sentiment is passed into the subject if a positive effect is expressed upon the object and expand upon the concept by also passing the reversed object’s sentiment into the subject if a negative effect is expressed.

Summarizing, even though we do not explicitly address the tasks of sentiment analysis and opinion mining, their resources in the form of lexicons and sentiment aggregation methods are relevant for our argument analysis in chapter 4.

Argumentative Relation Classification

Argumentative Relation Classification is a subtask of argument mining (Palau and Moens, 2009; Peldszus and Stede, 2013; Lippi and Torroni, 2016; Stede and Schneider, 2018). While there is some disagreement about the subtasks and their names, the core idea of argument mining is to detect arguments in a text, identify the conclusion and the premises, and determine the relations between arguments (attack or support, see section 2.1.1). Especially this last step, which we refer to as argumentative relation classification, is related to stance detection: Two arguments having the same stance towards a common topic are unlikely to attack each other, while if two arguments attack each other it is likely that they have a different stance towards the topic in question. More concretely, if the topic is formulated as a claim (e.g., *We should invest in renewable energies*), then classifying an argument’s stance towards the topic is essentially the same task as classifying its argumentative relation to the topic. We address such a task in chapter 3.

For argumentative relation classification, it is common to consider discourse markers and contextual features (Stab and Gurevych, 2014; Persing and Ng, 2016; Nguyen and Litman, 2016; Stab and Gurevych, 2017a). While discourse markers and such are certainly helpful for the classification task, we do not consider them particularly useful for understanding and explaining the arguments. Thus, in chapter 3, we decided to ignore such features even for the datasets where they would be available. In this regard, our work is similar to Menini and Tonelli (2016) and Hou and Jochim (2017) who also use Debatepedia to classify the relations between arguments. The relations they refer to are called *agreement* or *disagreement* and both terms are defined via the stances (having the same stance = agree, having different stances = disagree). Our work presented in chapter 3 expands upon these works by studying features based on external knowledge sources.

2.2.3 Relation Extraction and Knowledge Graph Generation

⁷ For both StArCon (chapter 4) and our Effect Relation Extractor (EREx) in chapter 5, we extract relations from text which we call effect relations. Our effect relation extraction is based on effect words, which indicate an effect their subject expresses on their object, and dependency parsing for identifying these subjects and objects. In StArCon, the effect relations are part of the effect triple. Concerning EREx, we then use the effect relations to build the effect graph.

Conceptually, our work is very similar to Al-Khatib et al. (2020) who also extract effect relations from argumentative texts and propose to use them to build a KG. Such a graph is then used by Al Khatib et al. (2021) where it serves as background knowledge to support neural argument generation, and by Yuan et al. (2021) who try to identify the correct response to an argument among five possible options. However, in terms of methodology, there are only little similarities to our approach. While StArCon and EREx are completely unsupervised, Al-Khatib et al. (2020) divide the relation extraction task into several subtasks for which they train specific classifiers, with one exception: For identifying the effect relation’s subject and object, they use the supervised OpenIE model of Stanovsky et al. (2018).

OpenIE (Open Information Extraction) is the task to extract relationships between entities from text. As such, it is relevant for both our effect triple and relation extraction

⁷This subsection is adapted from Kobbe, J., Hulpus, I., and Stuckenschmidt, H. (2023). Effect graph: Effect relation extraction for explanation generation. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 116–127, Toronto, Canada. Association for Computational Linguistics.

in chapters 4 and 5. In contrast to conventional information extraction, in OpenIE, the relationships are not predefined (Etzioni et al., 2008). However, OpenIE can also be applied for relation extraction with domain specific relations by performing *Relation Mapping* (Soderland et al., 2010). While Soderland et al. (2010) propose a supervised approach, in our use case, we consider it sufficient to filter and map the relations using an effect lexicon. Similarly to Corro and Gemulla (2013), Angeli et al. (2015), Gashteovski et al. (2017), we base our relation extraction on dependency parsing (see section 2.1.2). In contrast to these works, our effect triple and relation extraction approaches evolve around effect verbs specifically. This allows us to use only a small set of manually defined patterns, while still achieving comparable or better results than an OpenIE approach with relation mapping (see chapter 5.3.2).

Similar to our effect graph which we build from effect relations, Martinez-Rodriguez et al. (2018) use ClausIE (Corro and Gemulla, 2013) for extracting relations in order to build an OpenIE-based KG. Before applying OpenIE, they extract entities and link them to existing KGs. In our case, we experiment in EREx with both using only entities which can be linked to Wikipedia pages and not requiring any linking. Further, they annotate NPs and expand the extracted entities to encompass the complete NP. Similarly, in EREx we only consider NPs as entities.

Causal relations (Davidson, 1967) are conceptually similar to effect relations. Other than in effect relations, *A*’s effect on *B*, if they are in a causal relation, is clearly defined as *A* being the cause for *B* (and thus always positive). Girju and Moldovan (2002), Girju (2003) introduced the task of automatically extracting causal relations from text, and it has been a matter of research since then (Yang et al., 2022). Also for causal relations, there exists research on using them for building a KG. Heindorf et al. (2020) bootstrap dependency parse patterns to extract claimed causal relations from text. While their method to start with a small, very accurate seed set of patterns and to extend it consecutively is very appealing, we find it to be rather difficult to apply on our approach: Their patterns involve very concrete words that all trigger causal relations while we choose to keep our patterns general in order to apply to a large set of different effect words. Also like us, Heindorf et al. (2020) do not fact check their extractions, but emphasize that they merely collect claimed causal relations. This is also reflected by the wording of their label for effect relations, “may cause”, which we adopt to (*may*) *affect* for our own crowd annotation study in chapter 5.3.2.

2.2.4 Moral Foundation Classification

⁸ For classifying MFs (see section 2.1.4), there exist two main directions: dictionary-based and machine learning-based. In chapter 6, we experiment with both in order to classify MFs in argumentative texts.

Dictionary-based approaches

The first version of the MFD was presented by Graham et al. (2009) and has been used for a content analysis of christian sermons held in liberal and conservative congregations. Each of the five MFs (see section 2.1.4) has been further split into a vice and a virtue subcategory, reflecting the positive and negative ends of each dimension. Examples are *peace**, *protect**, *compassion** for *care_{virtue}* and *suffer**, *crush**, *killer** for *care_{vice}*. The MFD includes, on average, 32 words per moral subcategory. Frimer et al. (2019) present a new version of the MFD with more entries per MF subcategory, selected according to prototypicality estimates for each MF, based on cosine similarity for Word2Vec embeddings for each item in the dictionary. While the authors admit that the construct validity of the MFD 2.0 is not better than for the original MFD, they recommend the use of the MFD 2.0 due to its improved coverage. Rezapour et al. (2019) use WordNet to increase the original size of the dictionary to over 4,600 lexical items, using a quality controlled, human in the loop process. Similarly, Araque et al. (2020) use WordNet synsets to expand the MFD, additionally to expanding it with values for valence and arousal. However, despite using WordNet, the lexicons of both Rezapour et al. (2019) and Araque et al. (2020) are still on a word level. In contrast, we create and expand a word sense disambiguated version of the MFD.

The MFD has been used in several studies in the political and social sciences, psychology, and related fields (Takikawa and Sakamoto, 2017; Matsuo et al., 2018; Lewis, 2019). Dictionary-based approaches to measuring moral values in text, however, have severe shortcomings. They can neither account for unknown words or the different meanings a word can take, nor do they consider that shifter words and negation can change the polarity of an expression. In addition, we expect that moral vocabulary might vary considerably, depending on the speaker’s age and other geopolitical, social, and cultural variables. Garten et al. (2016) address the coverage problem of dictionary-based approaches by replacing the terms in the MFD with their averaged vector repre-

⁸This subsection is adapted from Kobbe, J., Rehbein, I., Hulpuş, I., and Stuckenschmidt, H. (2020b). Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

sentations in distributional space. They show that predicting moral foundations based on the cosine similarity of the words in a text to the distributional representations outperforms a naive method that predicts MFs based on word counts.

Machine learning-based approaches

Recent work has applied the framework of MFT to research questions in the social and political sciences (Fulgoni et al., 2016; Johnson and Goldwasser, 2018; Rezapour et al., 2019; Xie et al., 2019), replacing dictionary-based counts with more sophisticated methods. Johnson and Goldwasser (2018) model moral framing in politicians’ tweets, using probabilistic soft logic (Bach et al., 2013). Lin et al. (2018) improve the prediction of moral foundations by acquiring additional background knowledge from Wikipedia, using information extraction techniques such as entity linking and cross-document knowledge propagation. Xie et al. (2019) study the change of moral sentiment in longitudinal data, presenting a parameter-free model that predicts moral sentiment on three different levels: (i) moral relevance, (ii) moral polarity, and (iii) the ten moral subclasses of the MFD encoding the virtue/vice dimension for each MF. Finally, Rezapour et al. (2019) show that using dictionary counts for moral sentiment as features in a supervised classification setup can improve the results for stance detection.

2.2.5 Contributions

Summarized, our methodological contributions to the related work are as follows: We propose a method to reconstruct enthymemes of arguments from consequences specifically. We generate knowledge graph path based features for a sentence pair classification task and evaluate their impact on stance classification. Further, we introduce an unsupervised method to specifically detect stances for arguments from consequences (StArCon). While we do not outperform state-of-the-art stance detection methods, we consider our method to be valuable because it is efficient, explainable and because it provides effect triples extracted from the argument. Similarly to Al-Khatib et al. (2020), with EREx, we propose a method to extract effect relations from text and use it to generate a KG. Our method, again, is unsupervised and explainable, while, aside the focus on precision rather than recall, yielding comparable or slightly better results than the supervised method of Al-Khatib et al. (2020). Further, we also evaluate the resulting effect graph as a whole and its use for generating structured explanations for arguments from consequences which we propose. Lastly, we expand upon lexicon based MF clas-

sification by sense-disambiguating the MFD. We specifically evaluate MF classifiers on arguments and examine the usage of MFs in argumentative texts.

We also make contributions in the form of resources by publishing code, various expert- or crowd-annotated datasets and the effect graph. An overview of our published resources is provided in appendix A.

Chapter 3

Preliminary Work: External Knowledge for Argumentative Relation Classification

¹ The main intuition of this chapter is that for understanding an argument, the reader is often required to have specific background knowledge which is not spelled out in the argument itself, but can be found in KGs. Consider the following two statements:

Example 7. *The idea of lifelong marriage is outdated.*

Example 8. *Individuals should feel free to seek divorce.*

The first statement expresses a negative opinion towards *lifelong marriage*, and the second statement expresses a positive opinion towards *divorce*. At the same time, the two statements can be seen to *support* each other. In textual discourse, this relationship is often indicated with discourse markers like *because*, *therefore*, *thus* (i.e., Individuals should feel free to seek divorce, *thus* the idea of lifelong marriage is outdated.). Similarly, *attack* relations are frequently marked with discourse markers like *however*, *but*. Although the two examples have no words in common and do not include discourse markers, a human can easily determine the *support* relation between them. Essentially, this involves understanding the opposing relation between *lifelong marriage* and *divorce*.

¹This chapter is adapted from Kobbe, J., Opitz, J., Becker, M., Hulpuş, I., Stuckenschmidt, H., and Frank, A. (2019). Exploiting Background Knowledge for Argumentative Relation Classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASISs)*, pages 8:1–8:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

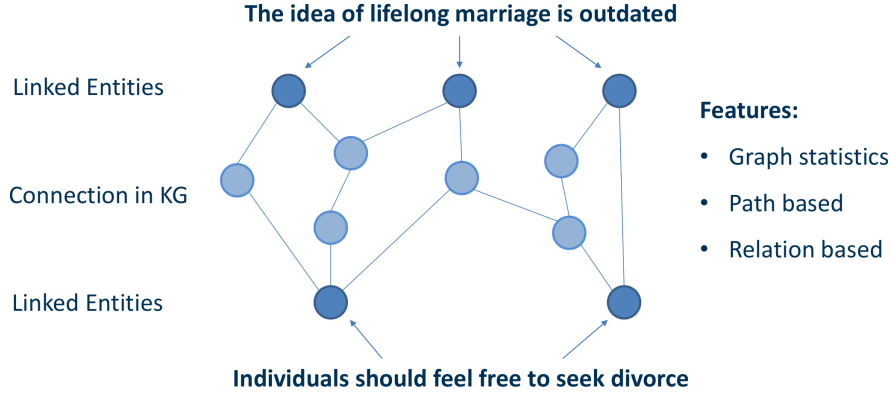


Figure 3.1: KG based features for sentence pair classification.

While accessing such knowledge is seamless for humans, it is much more challenging for machines. Traditional machine learning systems for argument analysis (e.g., Stab and Gurevych, 2017a; Afantenos et al., 2018) mainly rely on the exploitation of linguistic markers such as adverbials, discourse connectors or punctuation and largely ignore background knowledge and common sense reasoning as evidences for classifying argumentative relations.

Our main hypothesis is that combining text based features, like sentence embeddings, sentiment, and negation, with features incorporating background knowledge can improve the results for argumentative relation classification. Specifically, we are interested in a classification setup that is agnostic of the contextual surface features such as discourse markers and position in discourse, and that restricts classification to the analysis of two argumentative statements combined with the background knowledge that connects them. Because of this focus on local argumentative relation classification, our work is not directly comparable to prior work which proposes global, i.e., contextually aware classifiers for this task (Stab and Gurevych, 2014; Nguyen and Litman, 2016; Peldszus and Stede, 2015).

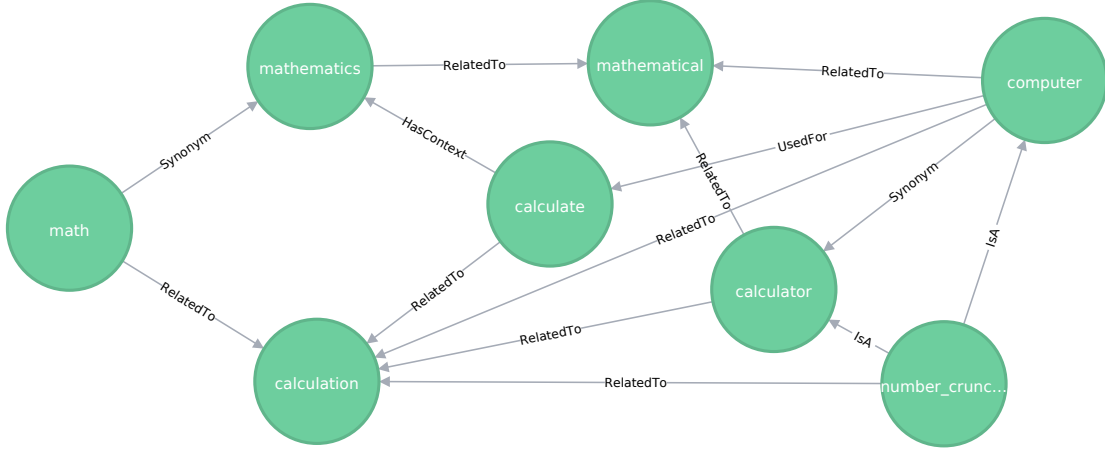
For extracting knowledge based features, we choose a straightforward approach which is illustrated in figure 3.1. We first link the entities in the two statements to a KG. In the first statement of the example provided in figure 3.1, depending on the KG, this could be *idea*, *lifelong*, *marriage* and *outdated*. The entities are then connected to those of the other statements in the KG of choice. We construct features based on graph statistics, as well as the occurring paths and relations. In this particular example, we expect especially the knowledge about the relation between *marriage* and *divorce* to be valuable. In ConceptNet, the two entities have direct edges indicating relatedness and antonymy.

In the remainder of this chapter, we present the concrete KG based features for sentence pair classification tasks in section 3.1. We propose a siamese neural network for sentence pair classification which can be extended with feature vectors encoding background knowledge in section 3.2. We evaluate the model on different argumentative datasets and using different sources of background knowledge to examine its impact on argumentative relation classification in section 3.3. Lastly, we conclude the chapter in section 3.4.

3.1 Knowledge Graph Features

For exploiting background knowledge, we design features based on two knowledge graphs: ConceptNet and DBpedia (see chapter 2.1.3). We expect ConceptNet to contain valuable information about common sense knowledge while DBpedia captures encyclopedic knowledge. The core idea is to connect two argumentative statements via relations in the KGs and to use the relation types and the extracted paths as features. The intuition is that certain types of paths or relations, e.g., the *Antonym* relation in ConceptNet, occur more often in disagreeing and therefore attacking pairs of statements than in supporting ones and vice versa.

Given two statements, we first proceed to link them to the external KGs. Section 3.3.2 provides the entity linking details. Once the two statements are linked, we represent them as sets A and B of their linked entities. We then pair all the elements in A to those in B . For each such pair $(x, y) \in A \times B, x \neq y$, we extract all the undirected paths from x to y up to length three within the KG. Figure 3.2 shows a graph consisting of such paths extracted from ConceptNet. As one can see in the graph, each path consists of nodes connected by directed edges labeled with *relation types*. As mentioned above, we assume that those relation types contain valuable information. For that reason, we design two kinds of features that rely on them: First, we check how often a certain relation type occurs along all paths between all pairs $(x, y) \in A \times B, x \neq y$ and divide that number by the total count of edges. This way, each relation type is a numerical feature on its own and all those features together sum up to 1. Second, we specifically exploit the paths. Since there are too many potential paths to create one feature per path, we group them via patterns. Each pattern is a multiset of relation types. For example, given the pattern $[Synonym, RelatedTo, RelatedTo]$, the graph in figure 3.2

Figure 3.2: Connection between *math* and *computer* in ConceptNet.

contains two paths between *math* and *computer* that instantiate this pattern:

$$\begin{aligned}
 & \text{math} \xrightarrow{\text{Synonym}} \text{mathematics} \xrightarrow{\text{RelatedTo}} \text{mathematical} \xleftarrow{\text{RelatedTo}} \text{computer} \\
 & \text{math} \xrightarrow{\text{RelatedTo}} \text{calculation} \xleftarrow{\text{RelatedTo}} \text{calculator} \xleftarrow{\text{Synonym}} \text{computer}
 \end{aligned}$$

Each such path pattern corresponds to a numerical feature whose value is the number of its instantiations divided by the total number of paths. As some of the relation type based and path based features described above occur only rarely, we only use those features that occur in at least one percent of all the instances in the training data.

Besides exploiting the relation types and paths, we also hypothesize that the length and number of paths are useful for classification, as they provide an indication to the relatedness of *A* and *B* (Hulpuş et al., 2015). To account for this, we additionally compute

- (i) a feature representing the total number of paths, divided by $|A| \cdot |B|$;
- (ii) three features representing the number of paths of a certain length i ($i \in \{1, 2, 3\}$), divided by the total number of paths;
- (iii) a feature representing the total number of identical entities in *A* and *B*, divided by $|A| \cdot |B|$;
- (iv) a feature representing the number of all the different nodes along all paths, divided by $|A| \cdot |B|$.

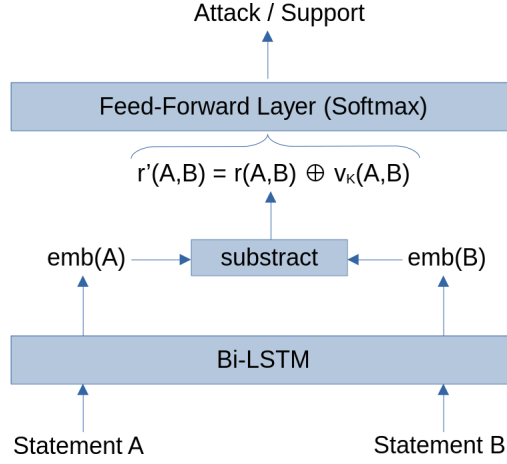


Figure 3.3: Architecture of the Siamese neural argumentative relation classifier.

3.2 Neural Network Model

² We design a Siamese neural network model for detecting an argument’s stance towards a topic, which is also expressed as a sentence. The architecture of the model is displayed in figure 3.3. It consists of one Bi-LSTM (Hochreiter and Schmidhuber, 1997), which is used to embed two argumentative statements A and B into a common vector space. More precisely, sequences of word embeddings³, $(e(w_1^A), \dots, e(w_n^A))$ and $(e(w_1^B), \dots, e(w_m^B))$ are fed through the Bi-LSTM to induce representations $emb(A)$, $emb(B) \in \mathbb{R}^{2h}$, where h is the number of the two LSTM’s hidden units (we concatenate the last states of the forward and backward pass of each LSTM). Based on the argument representations $emb(A)$ and $emb(B)$, we then compute a representation for the relation holding between these statements by computing the difference vector between their representations $emb(A)$ and $emb(B)$:

$$r(A, B) = emb(B) - emb(A) \quad (3.1)$$

The obtained representation for the relation can be further enriched by adding, e.g., features extracted from an external knowledge base that represent relevant information about knowledge relation paths connecting entities mentioned in the two argumentative statements (see section 3.1). The vector $v_K(A, B)$ that encodes such knowledge features is concatenated to the argument relation vector $r(A, B)$ to yield the extended vector

²The model presented in this subsection was proposed and implemented by Juri Opitz.

³We use pre-trained 300d Glove vectors (Pennington et al., 2014).

representation $r'(A, B)$ of the argumentative relation:

$$r'(A, B) = r(A, B) \oplus v_K(A, B) \quad (3.2)$$

$x \oplus y$ thereby denotes concatenation of vectors x, y . This final relation representation is further processed by a fully connected feed-forward layer with two output units and softmax-activations for providing the *support* and *attack* probabilities. We denote the neural network model by $NN+x$ whereby x depends on the source of the vector $v_K(A, B)$, or by NN for $v_K(A, B) = (0, \dots, 0)$.

3.3 Experiments

We conduct experiments on three argumentative data sets from different domains, which will be described in the following section. Because we want the models to focus on the background knowledge involved in the argumentation, we consider only the argumentative statements without their context and position. This increases the difficulty of the classification task as models are prevented from exploiting contextual and positional features, but we expect it to force the models to focus on the semantic relation between the two statements which we are interested in and which we ultimately aim to explain.

3.3.1 Data

Essays The student essays consist of 90 persuasive essays in the English language. The essays were selected from *essayforum*⁴ and annotated by Stab and Gurevych (2014). While also containing stance annotations, we use the dataset for argumentative relation classification. The corpus contains 1473 annotated argumentative relations. 1312 were labeled as *support* and the remaining 161 were labeled as *attack* relations. We apply the same split between training and test data as Stab and Gurevych (2014) and Nguyen and Litman (2016). While we make use of pairs of attacking and supporting statements, we dismiss all other information about the position and context and the annotated argumentative components and stances.

Microtexts This corpus consists of 112 short argumentative texts (Peldszus and Stede, 2016). The corpus was created in German and has been translated to English. We use only the English version. The corpus is annotated with argumentation graphs where the

⁴<https://essayforum.com/>

	Debatepedia	Microtexts	Essays
Total number of relations	14,441	308	1,473
Number of attack relations	7,184	84	161
Number of support relations	7,257	224	1,312

Table 3.1: Data statistics for the different experimental datasets.

nodes are argumentative units and the edges are argumentative functions. We collect pairs of attacking and supporting argumentative units to address argumentative relation classification. Therefore, we consider only direct connections between two argumentative units that are labeled as *support* or *rebut*. We deliberately ignore the *undercut* function as an undercut is an attack on the argumentative relation between two argumentative units. This way, we extract 308 argumentative relations whereof 224 are support and 84 are attack relations. To achieve a proper split between training and testing data, we use all the Microtexts about *public broadcasting fees on demand*, *school uniforms*, *increase weight of BA thesis in final grade* and *charge tuition fees* for testing and all the others for training.

Debatepedia Debatepedia.org was a website where users could contribute to debates on specific topics, which unfortunately is not accessible any longer. Most debates consist of a title, a topic that is formulated as a polar question (e.g. *Should the legal age for drinking alcohol be lowered?*), subtopics and arguments that are either in favor of or against the topic. We crawled the Debatepedia website and extracted all arguments with a valid URL. In many arguments, the argument’s claim is highlighted, so we used this feature to identify the claims, and removed the arguments that did not have any highlighted text. This resulted in 573 debates. We generate the pairs of statements by pairing the topic of the debate to the claim, essentially addressing stance detection. If the argument is in favor of the topic, then its claim *supports* the topic, else it *attacks* the topic. This way, we generate a large corpus containing 14441 pairs of statements whereof 7257 are in support and 7184 are in attack relations. We arbitrarily chose 114 (20%) out of the 573 debates for testing and use the rest for training.⁵

Table 3.1 shows the size and class distribution for all three datasets.

⁵The corpus and data split are available at <https://madata.bib.uni-mannheim.de/324/>.

3.3.2 Knowledge Graphs

DBpedia (DB) DBpedia contains information from Wikipedia in a structured way. For this particular usecase, we included the following datasets in English version in addition to the DBpedia Ontology (Version 2016-10): *article categories*, *category labels*, *instance types*, *labels*, *mapping-based objects* and *SKOS categories*. To achieve less meaningless paths, we excluded all the resources whose URI starts with *Category:Lists_of*, *List_of*, *Glossary_of*, *Category:Glossaries_of*, *Images_of*, *Category:Indexes_of*, *Category:Outlines_of*, *Category:Draft-Class*, *Category:Wikipedia* as well as the resource *owl:Thing*. For linking tokens in the argumentative units to entities in DBpedia, we use DBpedia Spotlight⁶ with a minimum confidence of 0.3 and support of 1.

ConceptNet (CN) ConceptNet is expected to contain commonsense knowledge (see chapter 2.1.3). We deleted all self-loops as they don't contain any valuable information. Linking of tokens to ConceptNet is done in a straightforward way: We split the statement into maximum-length sequences of words that can be mapped to concepts. If a concept consists only of stop words or has a degree of less than three, it is dismissed.⁷ This way, unconnected and only weakly connected concepts are avoided. If a concept consists of a single word, we use Stanford CoreNLP (Manning et al., 2014) to find out whether this is an adjective, noun or verb, in order to link it to the appropriate concept in ConceptNet, if possible.

3.3.3 Baseline

We train a linear classifier with replicated linguistic features, which we denote as *Ling*, and also experiment with using the features to augment our *NN*. As *Ling* features we use the sentiment of both argumentative units as features, as described in Menini and Tonelli (2016). We simplified the negation features of Menini and Tonelli (2016) and use Stanford CoreNLP (Manning et al., 2014) to only recognize whether there occurs negation in an argumentative unit. From Stab and Gurevych (2014), we adopted the structural features which contain token and punctuation statistics and two features indicating whether a modal verb occurs. Additionally, we use each pair of words, one from each statement, as a binary feature. We only included pairs that do not contain stopwords and occurred in at least one percent of all the training instances.

⁶<https://www.dbpedia-spotlight.org/>

⁷We use the default stopwords from <https://www.ranks.nl/stopwords> including *can*.

3.3.4 NN Model Optimization and Configurations

Optimization For parameter optimization, we randomly split off 200 examples from the training data of Debatepedia and Essays and 100 examples from the smaller Microtexts data. Let the training data be defined as $D = \{(x_i, y_i)\}_{i=1}^N$, where x_i consists of a source and target statement and $y_i \in \{0, 1\}^2$ is the one-hot vector corresponding to the two relation classes: (*support*, *attack*). Let, for any datum indicated by i , $p_{i,s}$ be the *support*-probability assigned by our model and $p_{i,a}$ the *attack*-probability. Using stochastic mini batch gradient descent (batch size: 32) with Adam (Kingma and Ba, 2014), we minimize the categorical cross entropy loss over the training data, H , computed as in Equation 3.3:

$$H = -\frac{1}{N} \sum_{i=1}^N (y_{i,s} \cdot \log p_{i,s} + y_{i,a} \cdot \log p_{i,a}), \quad (3.3)$$

where $y_{i,s} = 1$ if observation i is classified as *support* and 0 otherwise (and similarly $y_{i,a} = 1$ if observation i is classified as *attack* and 0 otherwise). We optimize all parameters of the model except the word embeddings.

Configurations Building on our basic Siamese model (NN), we inject (i) the graph features derived from ConceptNet ($NN+CN$); (ii) the same features but derived from DBpedia ($NN+DB$); (iii) a concatenation of both ($NN+DB+CN$). For comparison purposes, we also run experiments using only the feature vector derived from the knowledge base. This is achieved by basing the classification only on this feature vector (obtained from DBpedia (DB), ConceptNet (CN) or DBpedia and ConceptNet ($DB+CN$)), ignoring and leaving out the embedded relation. Instead of concatenating knowledge features to our Siamese relation classification model, we also perform experiments where we concatenate the linguistic feature vector to the argument relation embedding ($NN+Ling$). Our full-feature argumentative relation classification model is $NN+Ling+CN+DB$.

3.3.5 Results

Table 3.2 presents the F1 scores that our evaluated models obtain on all three datasets. Our first observation is that using only the KG based features yields very poor results. Further, NN on its own does outperform all baselines in terms of macro F1. With respect to the two targeted argumentative relation classes, *attack* relations are more challenging to capture in the Microtexts and Essays datasets, because of their very low frequency in the data (see Table 3.1).

	F1 scores								
	Debatepedia			Microtexts			Essays		
	support	attack	macro	support	attack	macro	support	attack	macro
random	50.2 \pm 1	50.1 \pm 1	50.2 \pm 1	73.0 \pm 5	27.8 \pm 11	50.4 \pm 8	89.2 \pm 1	10.5 \pm 4	49.8 \pm 3
majority	66.3	0.0	33.2	82.1	0.0	41.1	94.9	0.0	47.5
<i>Ling</i>	61.4	49.8	55.6	73.3	42.9	58.1	94.9	0.0	47.5
<i>DB</i>	43.7	56.8	50.2	81.1	0.0	40.5	94.8	0.0	47.4
<i>CN</i>	45.6	55.1	50.3	65.9	31.8	48.9	94.9	0.0	47.5
<i>DB+CN</i>	46.4	55.3	50.8	82.1	0.0	41.1	94.9	0.0	47.5
<i>NN+Ling</i>	58.1	55.7	56.9	77.7	35.2	66.7	92.7	20.7	56.7
<i>NN</i>	58.6	57.6	58.1	74.2	46.5	60.3	78.7	17.1	47.9
<i>NN+DB</i>	56.8	59.7	58.2	77.4	46.2	61.8	84.1	19.5	51.8
<i>NN+CN</i>	60.3	56.8	58.6	83.5	41.4	62.4	86.5	20.2	53.3
<i>NN+DB+CN</i>	58.6	57.6	58.1	81.2	38.7	59.9	88.0	16.3	52.1
<i>NN+Ling+CN+DB</i>	58.6	56.2	57.4	82.5	51.4	67.0	91.2	25.7	58.7

Table 3.2: Results over different systems and data sets.

Concerning the enhanced neural Siamese model (*NN+x*), the interpretation of the results is more unclear. For the two smaller datasets, *Microtexts* and *Essays*, adding *Ling* helps significantly more than adding KG based features, while adding both *Ling* and *CN+DB* is slightly better than *NN+Ling*. For our largest dataset, *Debatepedia*, however, it does not really seem to matter how *NN* is being enhanced. The results are pretty close to each other, and *NN* itself seems to perform reasonably well.

3.3.6 Analysis

To understand where the injection of background knowledge helps the most, we investigated the pairs of statements which were falsely classified by *NN* but correctly classified by *NN+CN* which had the best overall correction ratio. We rank these cases according to the margin $p_{NN+CN}(c) - p_{NN}(c)$, where $p(c)$ is the probability of the correct class. Four cases with large margins are displayed in table 3.3.

In the first example, *NN* mislabeled the relation as a *support* relation, assigning the attack relation a low probability. In contrast, the knowledge augmented model predicted the correct label confidently. For properly resolving example 1, it is somewhat surprising that the knowledge augmented model performs better than *NN*, since both statements express a stance towards the same target, namely *prohibition*. Background knowledge might still be needed in order to detect these stances, but we do not see a reason for why background knowledge *connecting* the two statements might be helpful.

In the second example, however, it is important to understand that *using technology or advanced facilities* is meant to be contrary to *investing much time in cooking*, which could plausibly be reflected in a connecting knowledge graph. Since *advanced facilities*

	Statement A (source)	Statement B (target)	y	Δ
1	prohibition has kept marijuana out of children's hands	prohibition does more harm than good	ATT	0.66
2	using technology or advanced facilities do not make food lose its nutrition and quality	investing much time in cooking food will guarantee nutrition as well as quality of food for their family	ATT	0.15
3	they will have a bad result in school	even people who are not interested in online game can still be negatively affected by using computer too much	SUP	0.84
4	Education and training are fundamental rights which the state, the society must provide	Tuition fees should not generally be charged by universities	SUP	0.38

Table 3.3: Examples from Microtext and Essays which were assigned a significantly higher probability for the correct label by the knowledge-augmented model ($NN+CN$) compared to our neural baseline model (NN).

are not an entity in ConceptNet, we investigate the paths between *technology* and *time*. There are 347 paths of length 3, so we display only the paths of length 2 in figure 3.4. But no matter the length, none of the paths is particularly helpful for understanding the relation between *technology* and *time*.

In example 3, again, we do not see a reason for why the connection of the few non-stopwords of statement A to statement B in a KG is particularly useful for finding the correct label. Basically, the key is to understand that statement A presents the negative effect which is postulated in statement B. Neither for understanding that statement A is an effect of using the computer too much nor that the effect is a negative one, we see how our knowledge features might help.

Example 4 is probably the most complex one. There are many relations between the statements which could be helpful, for instance to understand that universities provide education and that fundamental rights somehow contradict with charging fees. While the connecting graph indeed contains the edge *education* $\xrightarrow{\text{AtLocation}}$ *university*, there is no meaningful path between (fundamental) rights and fees.

3.4 Conclusion

We trained a neural model on different datasets and extended it with features from background knowledge to evaluate their usefulness in classifying relations between argu-

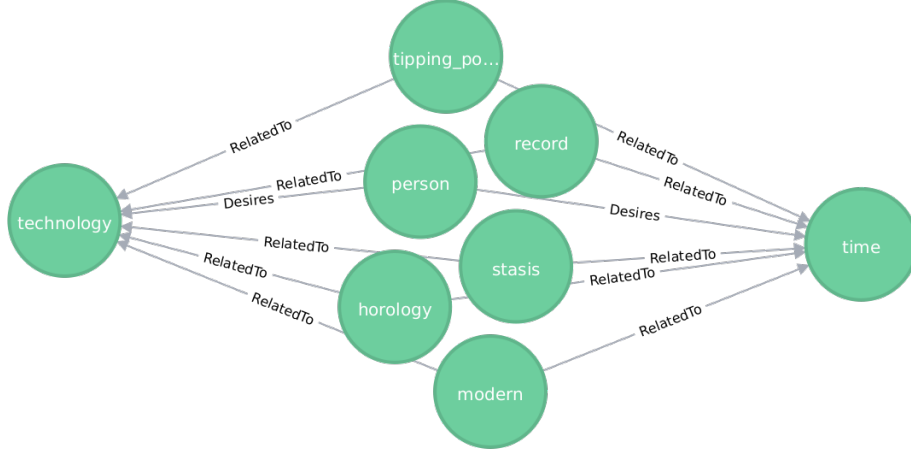


Figure 3.4: Paths of up to length 2 between *technology* and *time* in ConceptNet.

mentative statements. Though we observe some improvements especially on the smaller datasets, our results are not strong enough to draw a clear conclusion upon their impact. In the contrary: As indicated in our analysis of instances that were classified correctly potentially because of the insertion of background knowledge, we find the KGs we used to be lacking the relevant knowledge oftentimes. Arguably, one can achieve better results by incorporating the background knowledge in more sophisticated ways. Indeed, in a follow-up paper (Paul et al., 2020) we filter both the entities which we connect to the other statement and the resulting paths by using PPR. Along with some other changes, this leads to more consistent, but still small improvements. Overall, for both approaches, the quality and completeness of the KGs themselves set an upper limit of what is possible.

The work presented in this chapter motivated our further research in several ways. First, instead of trying to find a one-fits-all solution, we restrict ourselves specifically to arguments from consequences. This constraint allows us to deepen the analysis considerably by properly addressing the arguments’ reasoning. Second, we use background knowledge in a more targeted manner. For instance, in our example from the introduction, only the relation between *lifelong marriage* and *divorce* matters, but not the one between *idea* and *individuals* or *free*. Third, in chapter 5, we generate a KG containing specifically the type of knowledge we need and which we find to be lacking in existing KGs.

Chapter 4

Analyzing Arguments from Consequences

¹ Our first step towards analyzing arguments from consequences consists in reconstructing the scheme. This involves identifying (i) ACTION and CONSEQUENCE which make up the effect premise, (ii) CONSEQUENCE’s alignment which is part of the most often implicit judgment premise, and (iii) the conclusion whether or not ACTION should be brought about. The core intuition for our scheme reconstruction is that CONSEQUENCE is often expressed by a verb expressing a positive or negative effect that ACTION has on an object which we can assign a sentiment. This is in line with the templates of Reisert et al. (2018) (see chapter 2.1.1). Alternatively, CONSEQUENCE can consist of a sentiment word whose subject is ACTION and of its object. Below, we provide an example for each case.

Example 9. ACTION *increases the criminal rate*.

Example 10. ACTION *empowers women*.

In example 9, CONSEQUENCE is *increase of the criminal rate* which expresses a positive effect on an object with negative sentiment. In example 10, CONSEQUENCE is *empowerment of women* which expresses a positive effect in terms of sentiment on a sentiment-neutral object.

¹This chapter is adapted from Kobbe, J., Hulpuş, I., and Stuckenschmidt, H. (2020a). Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.

The corresponding data and source code are publicly available at <https://github.com/dwslab/StArCon>.

To evaluate our scheme reconstruction, we address the task of stance detection in this chapter. Since we base our stance detection on the reconstruction of the conclusion which in turn is based upon the premises’ reconstruction, stance detection provides an intuitive, though indirect way of quantifying the scheme reconstruction.

Stance detection is the task to decide whether a text is in favor of, against, or unrelated to a given topic. This problem is related to opinion mining, but while opinion mining focuses on the sentiment polarity explicitly expressed by a text, stance detection aims to determine the position that the text holds with respect to a topic that is generally more abstract and might not be mentioned in the text. As such, in stance detection, texts can transmit a negative sentiment or opinion, but be in favor of the targeted topic. In the following example, the argument expresses a negative opinion and sentiment, but its stance towards the topic is positive.

Example 11.

Topic *Criminalization of Holocaust denial*
Argument *Holocaust denial psychologically harms Holocaust survivors.*

The problem of stance detection in arguments has received growing attention from the scientific community, as shown by the survey of Küçük and Can (2020). Most traditional approaches tackle this problem by learning stance classification models for each debate topic. While this has the potential of achieving good results, new models need to be trained for each new topic of interest, generally entailing large annotation studies. More recently, transformer based approaches have shown to yield good results on the task (Schiller et al., 2020), which holds true even when applied on unseen topics or datasets as we will show in our evaluation. However, these models lack the explainability which we consider to be important for reaching our goal of explaining arguments.

While we admit that an explainable one-size-fits-all approach to stance detection is currently unfeasible, we take a different perspective: Rather than targeting topic-dependent models, we target arguments from consequences specifically. In most real-life arguments of this type, the consequences are expressed, but the interpretation that they are *good* or *bad*, as well as the conclusion, are most often implicit. The task of stance detection is then to determine if the argument is against or in favor of the topic.

For solving this task, aside reconstructing the argumentation scheme, we must relate the conclusion, or more specifically ACTION, to the topic. However, instead of first reconstructing the scheme and then relating ACTION to the topic, we first identify the

stance’s *target* in the argument by connecting it to the topic. In example 11, the target is *Holocaust denial*. In this case, ACTION matches this target, but it could also be a magnification or reduction of it (i.e., *approving Holocaust denial* or *criminalization of Holocaust denial*). After identifying ACTION, we identify the effect which ACTION expresses over its object. In example 11, the effect would be *(psychologically) harms* and the object *Holocaust survivors*. Together, the effect and the object form CONSEQUENCE whose polarity we identify by combining a sentiment and an effect lexicon. ACTION, effect and object form a triple which we denote as *effect triple* and which is at the core of our proposed method.

For the evaluation, we conducted an Amazon Mechanical Turk (mturk) study, in which we crowdsourced annotations for 1894 arguments extracted from Debatepedia. We compare our system’s performance to a sentiment analysis baseline and fine-tuned transformer based models. Our results are overall worse than the transformer based models’, though they are comparable or better in some settings. Aside from not needing annotated training data, we stress the advantage of our approach for providing human-understandable explanations to the results, and to provide, as a by-product, effect relations between concepts brought up in arguments.

The chapter is structured as follows: We propose a method to reconstruct the premises by extracting an effect triple in section 4.1. Further, we propose a method to use the effect triple for stance detection in section 4.2. In section 4.3 we describe the creation of our dataset. We evaluate our system in section 4.4 and compare it to state-of-the-art transformer-based models in section 4.5. We conclude in section 4.6.

4.1 Reconstructing the Premises: The Effect Triple

To demonstrate the reconstruction of the premises in detail, we look at the following example:

Example 12.

Topic *Medical marijuana dispensaries*

Argument *Legalizing medical marijuana does not increase use and abuse*

At the core of our approach resides what we call the **effect triple**. The effect triple is a triple of the form $\langle (Target, direction), (Predicate, effect), (Object, sentiment) \rangle$ which we abbreviate as follows: $\langle (T, dir), (P, eff), (O, sent) \rangle$. The (T, dir) pair represents the target T of the argument and if the argument refers to a magnification

Effect triple	Argumentation scheme	Example
(T, dir)	ACTION	Legalizing medical marijuana
P, O	CONSEQUENCE	no increase of abuse
$eff \cdot sent$	CONSEQUENCE's polarity	good

Table 4.1: Mapping between the effect triple and the argumentation scheme.

$(dir = 1)$ (e.g., *legalizing medical marijuana*), or a reduction ($dir = -1$) of the target (e.g., *banning medical marijuana*). The (P, eff) pair represents the predicate P that has T as the subject, together with the effect eff that it has over the object O . The effect can be positive ($eff = +1$) or negative ($eff = -1$). Lastly, the $(O, sent)$ pair represents the object over which T has the effect P . We expect the *sentiment* of an object to reflect whether it is generally regarded as *a good thing* ($sent = +1$) or *a bad thing* ($sent = -1$). In the example above, the effect triple is $\langle (\text{Medical Marijuana}, +1), (\text{not increase}, -1), (\text{abuse}, -1) \rangle$.

In terms of our notation of arguments from consequences, (T, dir) resp. *legalizing medical marijuana* corresponds to ACTION while CONSEQUENCE consists of both P and O : *no increase of abuse*. This reconstructs the effect premise. Whether CONSEQUENCE is considered good or bad (judgment premise) depends on eff and $sent$. Depending on the polarity, the conclusion can be deduced easily: If CONSEQUENCE is good, then ACTION should be brought about. If it is bad, then ACTION should not be brought about. Table 4.1 shows an overview of the mapping.

In the example, we consider the consequence of not increasing abuse to be positive. Strictly speaking, the argument does not present a positive consequence but the absence of a negative consequence. However, for solving stance detection, we treat *non-positive* consequences as negative and *non-negative* consequences as positive because generally such statements are used to attack respectively defend the target.

We now describe the lexicons and the method in more detail.

4.1.1 Lexicons

For determining dir , eff , and $sent$, we distinguish between the *sentiment* and the *effect* expressed by a word. For many words, the polarities of their sentiment and of their effect are the same (e.g., *kill*, *love*). Still, there are important exceptions, such as *reduce*, which has neutral sentiment but indicates a negative effect, or *conquer*, which has a slightly positive sentiment but indicates a negative effect on its object. Thus, we use an effect lexicon for determining dir and eff , and a sentiment lexicon for determining $sent$.

The Effect Lexicon To identify verbs and nominalized verbs that indicate effects on their direct objects, we extend the connotation frames (Rashkin et al., 2016). The connotation frames lexicon consists of a list of 947 verbs, manually annotated with values in the $[-1, 1]$ range, indicating if the verb implies a positive or negative effect over its object. We consider the entries with scores in the range $[-0.1, 0.1]$ as a neutral effect (e.g., *use*, *say*, *seem*), and we filter them out. We call the 845 remaining words in the lexicon **effect words**. We extend the list of effect words by adding all words in the same WordNet synset as the effect words, as long as there is no contradiction. A contradiction occurs when a new candidate effect word shares a synset with both a negative and a positive effect word. This way, we obtain 2508 effect words. We call this lexicon the Extended Connotation Frames lexicon (ECF). As ECF only contains verbs, we use it via the stems of the words, mainly to also get the effects of nominalized verbs. In our experiments, we compare the performance of this lexicon with +/-EffectWordNet (EWN) (Choi and Wiebe, 2014).

The Sentiment Lexicon In order to determine if the object of the effect is something *good* or *bad*, we combine several commonly used sentiment lexicons: (i) the subjectivity lexicon² of Wilson et al. (2005); (ii) the opinion lexicon of Hu and Liu (2004); and (iii) the sentiment lexicon of Toledo-Ronen et al. (2018) (uni- and bigrams, using a threshold of ± 0.2). The composed lexicon contains sentiment values in the range $[-1, 1]$.

4.1.2 Effect Triple Extraction

Target Identification As *target*, we denote the entity towards which the argument expresses a stance and which is directly related to the topic. The target can be ACTION or a part of ACTION. Oftentimes, such as in example 12, the target is explicitly mentioned in both the argument and the topic. However, the target can also be referenced differently in the argument and the topic which is why we denote the argument’s target by T_a and its mention in the topic by T_t . The latter is needed for addressing the stance detection task in section 4.2. To detect both T_a and T_t , we exploit their semantic relatedness. Thus, we identify T_a and T_t simultaneously by following three strategies. The use of the second and third strategies is conditioned on the previous strategies to have failed to identify a pair of targets. First, we look for a pair of nouns that are identical or

²We used an American English dictionary to correct orthographic mistakes resp. to add American English versions of British English words.

Pattern	Interpretation	Example
1 $P \xrightarrow{*} O$	P has object O	Insurance mandates violate the rights of employers. <div style="text-align: center;"> $\boxed{\text{dobj}} \uparrow$ </div>
2 $P \xrightarrow{\text{prep} ? \text{pobj}} O$	P has object O	The military industrial complex profits from escalation in Afg. <div style="text-align: center;"> $\boxed{\text{prep}} \uparrow \boxed{\text{pobj}} \uparrow$ </div>
3 $P \xrightarrow{\diamond} S$	P has subject S	Holocaust denial is inherently discriminatory and damaging. <div style="text-align: center;"> $\uparrow \boxed{\text{nsubj}}$ </div>
4 $X \xrightarrow{\dagger} M \wedge \text{sent}(M) \neq 0$	$\text{sent}(X) := \text{sent}(M)$	W/o more troops, Afg. will become terrorist haven <div style="text-align: center;"> $\uparrow \boxed{\text{amod}}$ </div>
5 $\text{NegP} \xrightarrow{\text{pobj}} X$	X is negated	Free speech without Fairness Doctrine can harm policy-making <div style="text-align: center;"> $\boxed{\text{pobj}} \uparrow$ </div>
6 $X \rightarrow \text{NegP} \wedge \nexists \text{NegP} \xrightarrow{\text{pobj}}$	X is negated	W/o more troops, Afg. will become terrorist haven <div style="text-align: center;"> $\uparrow \boxed{\text{nn}}$ </div>
7 $X \xrightarrow{\text{neg}}$	X is negated	Solar energy does not damage air quality. <div style="text-align: center;"> $\uparrow \boxed{\text{neg}}$ </div>

Table 4.2: Dependency graph patterns. $*$ $\in \{\text{dobj}, \text{nsubjpass}, \text{cobj}, \text{csbjpass}, \text{nmod}, \text{xcomp}\}$; $\diamond \in \{\text{nsubj}, \text{csbj}\}$; $\dagger \in \{\text{amod}, \text{nn}, \text{advmod}\}$; NegP stands for *negative preposition*.

have the same lemma. We use Stanford Core NLP (Manning et al., 2014) for POS tagging and lemmatizing. Second, we look for a pair consisting of an acronym (e.g., *ICC*) and a word sequence whose first letters form the acronym (e.g., *International Criminal Court*). Third, we look for pairs of nouns that are synonyms or antonyms according to *Thesaurus.plus*³.

Besides returning T_a and T_t , we also return a value $r = +1$ if the two targets have been found to be synonyms and $r = -1$ if they are antonyms. Thus, first and second strategies only return $r = 1$ while the third strategy returns 1 or -1 .

Target Direction Determination As described earlier, each target is accompanied by a *dir* value which indicates if the statement refers to a phenomenon of amplification or reduction of the target. We detect this by searching for a word whose object is the target by using patterns 1 and 2 shown in Table 4.2. The word is then looked-up in the effect lexicon. If a negative effect is found, then $\text{dir} = -1$, otherwise $\text{dir} = 1$. We call the word the *target effector*, or just *effector*. In the argument in example 12, the effector is *legalizing* and expresses an amplification of the target ($\text{dir} = 1$).

Detecting Predicates and Their Effects Effect words are commonly used in arguments from consequences to express a (potential) effect that the target has or might

³We use only the synonyms and antonyms shown at <https://thesaurus.plus/thesaurus/xxx> where xxx is a placeholder for concrete words.

have over another object. For example, in the argument in example 12, the effect word *increase* expresses a positive effect that the (amplified) target has over the objects *use*, *abuse*.

We detect this effect of the target by using pattern 3 to find a predicate whose subject is either the target or its effector, and by looking up this predicate in the effect lexicon. We thereby set eff to 1 or -1 , depending on if the effect is positive or negative. In our running example, the (P, eff) pair becomes $(not\ increase, -1)$ because of the negation, as we explain at the end of this section.

Telling good from bad The last effect triple component we detect is $(O, sent)$. To this end, we search the dependency graph for instantiations of patterns 1 or 2, where P is the predicate that has been detected to express the target’s effect. If such an object is found, we use the sentiment lexicon by first searching for the exact word and, if not available, for the word’s lemma. We set $sent$ to -1 if the word bears a negative sentiment or to 1 otherwise. In our example, the $(O, sent)$ pair becomes $(abuse, -1)$ because the word *use* is neutral per se.

The sentiment of a word is overwritten by the sentiment of its modifiers, as shown in pattern 4 in Table 4.2. In the provided example in the table, one can see that the modifier *terrorist* dominates the sentiment of the positive word *haven*. Consequently, both *terrorist haven* and *terrorist attack* are considered generally bad.

Negation We deal with negations for each effect triple component. We identify negations by looking for patterns 5, 6, and 7, as shown in Table 4.2. Patterns 5 and 6 make use of a manually created list of all negative English prepositions.⁴ The existence of a negation affecting the target, predicate, or object toggles the sign of the corresponding value (dir , eff or $sent$, respectively).

In the following, we detail how we use the effect triples for stance detection.

4.2 Stance Detection with Effect Triples

Inferring the Stance of the Claim Towards the Target In order to detect the stance towards the topic, we first infer the argument’s stance towards the target. For that purpose, we use the intuition that the stance is unfavorable when the text expresses negative

⁴Those are *except*, *less*, *minus*, *opposite*, *sans*, *unlike*, *versus*, *without*, *w/o*, *vice*, *instead (of)*, *lack*.

consequences of the target, and positive otherwise. Thus, we define that the stance towards the target is positive in exactly the following four cases:

- (i) the target's amplification implies a positive effect over something good:

$$dir = eff = sent = +1$$

- (ii) the target's amplification implies a negative effect over something bad:

$$dir = +1, eff = sent = -1$$

- (iii) the target's reduction implies a negative effect over something good:

$$dir = eff = -1, sent = +1$$

- (iv) the target's reduction implies a positive effect over something bad:

$$dir = +1, eff = -1, sent = +1$$

Hence, the stance is favorable towards the target if the multiplication of the three components' values is $+1$. Consequently, we define the argument's stance towards the target as

$$s_a = dir \cdot eff \cdot sent \quad (4.1)$$

and interpret $s_a = 1$ as *in favor* and $s_a = -1$ as *against*.

Inferring the Stance of the Claim Towards the Topic The steps above can be executed analogously for the argument and the topic. However, due to the nature of the text expressing the topic, we only aim to extract an effect triple from the argument. For the topic, we detect its target T_t and set the stance s_t to its corresponding dir value. To infer the argument's stance towards the topic, we need to consider the relation between T_a and T_t , i.e., the value of r as described in Section 4.1.2. We then define the final result of the analysis as

$$\Pi = s_a \cdot s_t \cdot r \quad (4.2)$$

Table 4.3 presents further examples of how our approach detects the stance of the argument towards the topic. As illustrated in the examples, the straightforward interpretability of the stance detection process can be easily used for producing human-readable explanations for the returned results. This is particularly relevant for helping users get more control over the process, particularly in light of subsequent applications on top of stance detection.

Alternative Strategies We denote the process in which all the previous steps are fulfilled and an effect triple is extracted as **TPO** (*target, predicate, object*). However, due

	<i>Argument</i>	<i>Topic</i>
	Porn watching may actually reduce rape rates	Pornography
T, dir P, eff $O, sent$	<i>Porn</i> , +1 <i>reduce</i> , -1 <i>rape rates</i> , -1	<i>Pornography</i> , +1
s	1	1
r	1	
Π	1 (In favor)	
	Holocaust denial psychologically harms Holocaust survivors	Criminalization of Holocaust denial
T, dir P, eff $O, sent$	<i>Holocaust denial</i> , 1 <i>harms</i> , -1 <i>survivors</i> , +1	<i>Holocaust denial</i> , -1
s	-1	-1
r	1	
Π	1 (In favor)	

Table 4.3: Worked out examples.

to a variety of reasons that we analyze in section 4.4.4, we might fail to extract a complete effect triple. This is the case if the object has no sentiment such as in *empower women*, or if the effect is expressed, for instance, by an adjective such as in *Holocaust denial is discriminatory*. For that reason, if we identify T_a and P but not O , we set eff to the sentiment polarity of P and $sent$ to +1 by default. We denote this strategy by **TP**.

Another potential situation is that the system detects (P, eff) and $(O, sent)$, but it can not relate them to T_a . One cause can be that we fail to identify T_a and T_t . If so, $dir = +1$ by default. Another cause can be that T_a is found, but we can not infer its relation to P . In this case, we consider that the identified target is the subject of P and set (T_a, dir) accordingly. We refer to this strategy as **PO**.

Lastly, if all above strategies fail to create an effect triple, we use a heuristic: if T_a and T_t were found, dir is set accordingly. Otherwise $dir = 1$ by default. For the remaining words in the statement, we check their sentiment score, still using pattern 4, toggling the sign if it is negated. The sum of the sentiment scores is then multiplied with dir . The stance is considered favorable or not depending on the sign of the result. We refer to this strategy as **Heuristic**.

4.3 Dataset Generation

To evaluate our approach, we need stance annotated topic-argument pairs, as well as annotations whether the topic-argument pair refers to a consequence or not.

4.3.1 Data Collection

To create such a corpus, we run a mturk crowdsourcing study, where we annotate argumentative claims and topics extracted from Debatepedia. Other than in our experiment in chapter 3, we now only use the 236 *Featured Debate Digest* articles as they are of higher quality. They contain more than 10,000 arguments labeled by their author as either pro or con the debate’s topic. Usually, the arguments start with a bolded, one-sentence summary, which serves as the argument’s claim. We exclusively use these claims and pair them to the debate’s topic. We exclude 16 debates whose topics contain *vs* or *or* (e.g. *Democrats vs. Republicans*), and 30 debates without a title question. To create a balanced dataset that covers a large variety of topics, we randomly selected 5 pro and 5 con arguments of each debate. If a debate contains less than 5 pro and 5 con arguments, we select the maximum equal number of pro and con arguments. We obtain 190 different topics and 1894 arguments.

4.3.2 Crowdsourcing Study

The annotation task consists of the debate’s topic, one of its arguments, and two questions. The first question is to select the stance of the argument towards the topic, out of the following choices: *in favor*, *against*, *neither* and *I don’t know*. Although we have the original arguments’ stances, this question helps us check how clear the argument is when taken out of the debate’s context. The second question is whether the argument refers to a consequence related to the topic, with possible answers *yes*, *no* and *I don’t know*. Each topic-argument pair was annotated by 10 annotators living in the US with a HIT approval rate greater than 98% and more than 10,000 approved HITs in total. Overall, 277 annotators worked on the task.

4.3.3 Agreement and Reliability

Table 4.4 shows the IAA per number of valid annotations, i.e., annotations that are not *I don’t know*. Since we have many annotators, Fleiss κ is particularly low on consequence

Valid Annotations	<i>Stance</i>			<i>Consequence</i>		
	rate	κ	κ'	rate	κ	κ'
6	.002	-.10	-.20	.001	-.17	-.1
7	.013	.11	.15	.008	.04	.10
8	.051	.24	.32	.036	.06	.24
9	.183	.34	.58	.207	.23	.44
10	.751	.52	.74	.748	.25	.58
Weight. Avg		.47	.68		.24	.53

Table 4.4: Fleiss' Kappa dependent on the number of valid annotations.

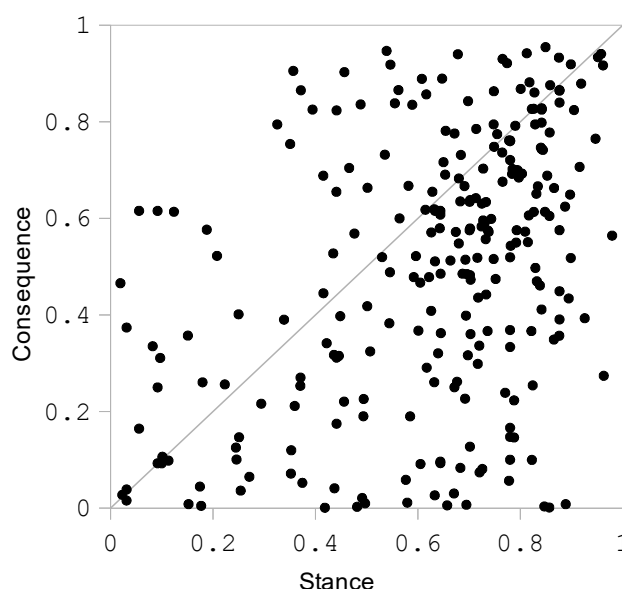


Figure 4.1: Reliability of annotators according to MACE: The higher the score, the more reliable the annotator is.

annotation, but still indicates higher agreement than random. To give an agreement estimate less sensitive to individual outliers, we also compute κ' as the Fleiss kappa between two *experts*, where each expert brings together half of the number of annotators and its annotation is decided by Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013).

Figure 4.1 shows the reliability of individual annotators. Although there is a weak correlation among the reliability of the two tasks (Pearson .41), some annotators are quite reliable in annotating stances, but highly unreliable in annotating consequences. This indicates that the latter task was unclear to some of the annotators. To understand why the annotators usually disagree, we investigated such instances and identified several possible reasons:

Complexity In the topic-argument pair *Criminalization of Holocaust denial – Danger of public accepting holocaust denial should be fought by logic*, both topic and argument have a negative stance towards *holocaust denial*, which suggests the label *in favor*. Still, by proposing a different solution than *criminalization*, the argument is *against* the topic.

Lack of Domain Knowledge Many arguments involve non-trivial background knowledge: *Israeli military assault in Gaza – Hamas was first to escalate conflict following end of ceasefire*.

Ambiguity Stances and consequences can be ambiguous: *2009 US economic stimulus – Stimulus risks being too small not too large*. A small stimulus is bad while an appropriate stimulus is good.

Ethical Judgement Different judgments on what is good and bad can lead to different stance labels: *Ban on human reproductive cloning – Cloning will involve the creation of children for predetermined roles*.

Lack of Conceptual Clarity Presumably, especially for the consequence annotation, there was too much room for interpretation regarding our task description. For example, in *Solar energy – Solar energy does not damage air quality*, three annotators stated that the argument does not “refer to or suggest a consequence related to the topic”, while the other seven stated that it does.

4.3.4 Final Dataset

To account for unreliable annotators, we compute the annotation result with MACE. As such, we find that for 81.36% of the annotated arguments, the stance label obtained via MACE is the same as the original stance label. By comparison, the majority vote matches 79.30% of the original stance labels. Since disagreements between the MACE annotation and the original stance might indicate that the argument’s stance is unclear outside the debate’s context, we exclude all such pairs from the dataset. For example, the original label of the pair *Is Wikipedia valuable? – Wikipedia is online and interactive, unlike other encyclopedias* is *con*, because, in its context, it was discussed whether Wikipedia is an encyclopedia or not. In contrast, understandably, the result of our annotation is *pro*. Since the original labels are only *pro* or *con*, all pairs that our study

<i>conseq</i>		<i>other</i>		<i>debate</i>		<i>wiki</i>	
pro	con	pro	con	pro	con	pro	con
376	446	370	310	746	756	1195	1199

Table 4.5: Class distributions.

determined as *neither* are removed. This filter resulted in a total of 1502 pairs, out of which 822 have been annotated to relate to consequences.

4.4 Evaluation

4.4.1 Data

We report results both on the 822 pairs that relate to consequences, denoted by *conseq*, and on the rest of the pairs, denoted by *other*, as well as on their union, denoted by *debate*.

For checking the performance of the systems on an independent dataset, we also use the claim stance dataset⁵ published by Bar-Haim et al. (2017a). This dataset contains 55 topics of *idebate*⁶ and 2394 manually collected claims from Wikipedia. We denote this dataset by *wiki*. As Bar-Haim et al. (2017a,b) do, when working with this dataset, we use only the topic’s target and not the entire topic to ensure comparability.

Table 4.5 shows the class distribution of the datasets.

4.4.2 Compared systems

We evaluate our system with the effect lexicon that we describe in section 4.1.1 (ECF), as well as with the EWN. We denote the system by Stance Detector for Arguments from Consequences (StArCon). For comparison, we implement two other approaches:

sent As a baseline, we use a system that simply sums up all the sentiment scores in the argument. For the *wiki* dataset, the sign is switched if the topic sentiment is negative.

BERT BERT (Devlin et al., 2018) has shown to outperform a series of alternative stance detection systems (Ghosh et al., 2019). We fine-tune BERT using the large,

⁵Available at https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml

⁶<https://idebate.org/>

	conseq				other				debate				wiki			
	pro	con	mac	acc	pro	con	mac	acc	pro	con	mac	acc	pro	con	mac	acc
<i>sent</i>	.62	.67	.65	.65	.64	.47	.56	.57	.63	.59	.61	.61	.61	.58	.60	.60
<i>BERT</i>	.65	.82	.74	.78	.73	.48	.60	.66	.63	.72	.67	.71	.72	.65	.68	.70
- <i>BERT std dev</i>	.33	.08	.20	.13	.06	.31	.17	.11	.32	.18	.21	.15	.07	.24	.15	.11
<i>StArCon</i>	.72	.74	.73	.73	.69	.56	.63	.64	.71	.67	.69	.69	.66	.63	.64	.64
<i>StArCon (EWN)</i>	.70	.72	.71	.71	.66	.53	.60	.61	.68	.64	.66	.66	.64	.61	.63	.63

Table 4.6: Experimental results. F1 scores per stance class (*pro* and *con*), macro-F1 (*mac*), and accuracy (*acc*). For BERT, we show the mean of the respective cross-validation results and their standard deviation.

uncased pre-trained weights.⁷ Just as Schiller et al. (2020), we set the number of epochs to 5 and the batch size to 16. The input are topic-argument pairs. We perform 10-fold cross-validation with a train-dev-test ratio of (70/20/10), ensuring that each topic exclusively occurs in one set.

4.4.3 Results and Discussion

The results that compare StArCon to BERT and the sentiment detection baseline are presented in Table 4.6. First, as expected, StArCon performs better on arguments related to consequences than on other arguments, with a macro-F1 difference of 9% between *conseq* and *other*. Further, StArCon with both lexicon settings consistently outperforms the *sent* baseline, but its macro-F1 score is outperformed by BERT on *conseq* and *wiki*, and its accuracy is outperformed by BERT on all datasets. This is not surprising, given that we use BERT pre-trained and then fine-tuned to our data. Interestingly, StArCon with ECF achieves better results than BERT in terms of macro F1 score on the arguments that are *not* related to consequences (*other*), and on the complete *debate* dataset. This indicates that our method can deal reasonably well with arguments that are not from consequences.

Concerning the two stance classes, with both lexicon settings, StArCon is better than BERT at predicting the *pro* class in arguments from consequences, but is outperformed on the *con* class. Another interesting result is that on *conseq*, StArCon has a quite similar performance on the *pro* and *con* classes with both lexicon settings. In contrast, BERT’s performance varies drastically, with a difference of approximately 17% in favor of the *con* class. BERT’s high variability is also indicated by the high standard deviation on the 10 folds. For comparison, we also computed the F1 macro standard deviation of StArCon with ECF when run on the same 10 folds, and the values lie between .03 on

⁷We worked with the original release: <https://github.com/google-research/bert>

	conseq		other		debate		wiki	
	<i>r</i>	F1	<i>r</i>	F1	<i>r</i>	F1	<i>r</i>	F1
<i>Total</i>	1	.73	1	.63	1	.69	1	.64
<i>Target found</i>	.82	.74	.76	.64	.80	.70	.53	.67
-Word/Lemma	.75	.74	.72	.64	.74	.70	.42	.67
-Acronym	.02	.80	.01	.89	.02	.83	.00	–
-Synonym/Antonym	.05	.69	.03	.50	.04	.64	.11	.66
<i>TPO/TP/PO</i>	.60	.76	.39	.64	.51	.72	.54	.67
-TPO	.23	.74	.05	.65	.15	.73	.07	.81
-TP	.21	.84	.18	.74	.20	.80	.10	.77
-PO	.16	.69	.16	.53	.16	.62	.36	.62
<i>Heuristic</i>	.40	.68	.61	.61	.49	.65	.46	.61

Table 4.7: Evaluation of the target identification and stance detection strategies; *r* denotes the rate of data instances.

debate and .07 on *conseq*. This indicates that our unsupervised approach is more robust with more predictable performance.

Concerning the two effect lexicons, StArCon performs consistently better when using ECF than when using EWN. Our analysis indicates that the high coverage of the EWN lexicon comes at the expense of accuracy. Therefore, in the following, we will only refer to StArCon using ECF.

Regarding the two datasets *debate* and *wiki*, BERT outperforms StArCon, with quite a high margin particularly on the *wiki* data. The accuracy that Bar-Haim et al. (2017a,b) report on the *wiki* data, when no context features are used, is .68 which is lower than BERT’s (.70) but higher than StArCon’s (.65 for evaluating on the dedicated test set). This is not surprising given that the data contains general arguments. Nevertheless, as our approach only targets a subclass of these arguments, the results are quite promising. Unfortunately, the system of Bar-Haim et al. (2017a,b) is proprietary and we could not evaluate it on our *conseq* data.

Table 4.7 provides further insights into our solution. First, on all Debatepedia based datasets, we find a target in more than .75 of the data instances, and overall, the results are slightly better when a target is found. Most of the targets are found by word similarity and the fewest by the acronym. The results obtained on the instances where the target was found by synonym/antonym relations are significantly lower than those obtained when the target was found with the other two strategies. This indicates that the approach is sensitive to semantic drift in target identification.

Overall, we identify a potential consequence (*TPO/TP/PO*) for .6 of the arguments

in *conseq*. While the results are quite good on all datasets when we detect a complete effect triple (*TPO*), they are overtaken by results of the *TP* cases. Together, the instances solved with *TPO* and *TP* strategies amount to .44 of the *conseq* dataset but to much lower on the other datasets (e.g., only .17 on the *wiki*). The performance on the *PO* cases is comparable to the performance on the *Heuristic* cases, and significantly lower than when *TPO* or *TP* could be applied. Depending on the dataset, the system needed to apply the *Heuristic* strategy on .4 to .61 of the instances. Thus, potentially, helping the system make sense of more of the arguments so that the number of times it needs to fallback to *PO* and *Heuristic* is reduced would significantly improve the results.

4.4.4 Error Analysis

To better understand the limitations of our approach, we analyzed the errors on the *conseq* data and found several reasons for wrong predictions:

Incomplete list of patterns Some arguments cannot be meaningfully analyzed with our current list of patterns. Potentially, one can extend this list with more complex patterns or automatically learn such patterns from data.

Conceptual errors We assume that positive effects on something negative result in something negative (e.g., *War in Iraq has helped terrorist recruitment.*). However, this is not always the case (e.g., *Privatizing social security helps the poor.*).

Finding the targets As shown in Table 4.7, we often fail to detect targets. For example, our target detection strategies fail on the argument-topic pair *Standardized tests ensure students learn essential information. – No Child Left Behind Act*. In this specific case, there is a hypernym relation between the topic and *Standardized tests*. Further, we found that our straightforward approach to identifying targets and the relations between them is one of the core reasons for our approach’s poorer performance on the *wiki* data compared to the *debate* data. The target finding strategy might be improved by leveraging additional semantic knowledge.

Missing / wrong lexicon entries For many words, we are missing an entry in our lexicons, or the entry exists but is questionable. For instance, in the sentiment lexicon, *Palestinian* is annotated with a negative sentiment. Also, sometimes the effect on the object seems to be mixed up with the word’s overall effect. For example, *solve* has

a positive effect on the object in both ECF and EWN lexicons, but arguably when a problem is *solved*, it undergoes a reduction (e.g. *Reforestation,[...] can help solve global warming*).

Ambiguity Some words have a positive or negative effect depending on the sense with which they are used (e.g., *push* vs. *push for*). In the effect lexicon, we have only one entry per word. In the EWN, there are multiple senses, but we always use the most probable effect. Word sense disambiguation is required for these cases, which is known to be very challenging for verbs. However, a potential solution could be to annotate VerbNet frames (Schuler, 2005) with effects.

Text parsing errors As our method relies on the output of the dependency parser, the Lemmatizer, the POS tagger, and the Stemmer, their errors naturally propagate.

4.5 Transformers for Stance Detection

Since the performance of the BERT model we compared StArCon to in section 4.4 is rather inconsistent and the original release being outdated, we perform a deeper analysis of the performance of transformer-based models for stance detection. Aside from training and testing on the same dataset, we also perform a cross-domain evaluation where we train and test on different datasets.

Models We compare the large (uncased) versions of BERT, RoBERTa, DeBERTa. We use the Huggingface implementation (Wolf et al., 2020). Again, we set the number of epochs to 5 and the batch size to 16 as proposed by Schiller et al. (2020).

Datasets The datasets we use for the evaluation are the same as described in section 4.4.1. Additionally, we use our full parse of the featured debates of debatepedia.org, denoted by *full-debate*. As our crowd study indicates, this dataset contains gold labels which may be considered wrong, but on the other hand it is considerably larger than the other datasets we use. It contains 10,446 arguments of which 5,261 are in favor and 5,185 are against their respective topic.

	conseq		other		debate		full-debate		wiki	
	avg	dev	avg	dev	avg	dev	avg	dev	avg	dev
BERT	82.7	5.6	66.3	6.3	80.2	3.1	75.5	3.3	73.6	8.1
RoBERTa	69.6	19.2	60.1	9.5	80.0	16.3	81.1	10.8	84.7	9.9
DeBERTa	91.7	3.0	85.6	7.8	91.4	2.5	86.4	3.1	90.0	3.3
StArCon	72.8	7.3	63.8	6.5	68.8	2.9	62.1	2.1	64.3	4.7

Table 4.8: In-domain evaluation for stance detection: average (avg) and standard deviation (dev) of the accuracy scores in percentages.

4.5.1 In-Domain Evaluation

Similarly to our evaluation in section 4.4, we perform a 10-fold cross-validation, ensuring that each topic exclusively occurs in either the train or the test set. Since we do not necessarily need a development dataset, we use the typical 90-10 train-test split this time. Thus, the training sets are larger than in our previous evaluation. For comparison, we also include StArCon as described in this chapter, using the same evaluation setting (dismissing the training data). Table 4.8 shows the results.

RoBERTa struggles on the small datasets (especially *conseq* and *other*), but outperforms BERT for the larger ones. On the other hand, DeBERTa consistently performs best. Especially when put into perspective with our crowd annotation study, DeBERTa’s results are impressive: In our random selection of 1894 arguments, the majority vote of the crowd workers matches 79.30% of the original stance labels. For the stance label obtained via MACE, it is 81.36%. DeBERTa’s average accuracy on the dataset we randomly selected the arguments from is 86.4%. Concluding, even though RoBERTa is sometimes worse than StArCon, modern BERT-like models are better suited for stance detection than StArCon.

4.5.2 Cross-Domain Evaluation

To assess how well the models generalize over different datasets, we train them on one of the subsets of *full-debate* and test them on *wiki*, or vice versa. Additionally, we train on *conseq* and test on *other* and vice versa, as these are the only non-overlapping subsets of *full-debate*. For easier reference, we also include StArCon’s results when tested on these datasets. Table 4.9 shows the results.

Overall, the results are worse than those in the in-domain setup, but only by a small amount. DeBERTa consistently outperforms the other models, while RoBERTa is better than BERT except, when trained on *other*. Generally the models perform better on *wiki*

BERT	<i>conseq</i>	<i>other</i>	<i>debate</i>	<i>full-debate</i>	<i>wiki</i>
<i>conseq</i>		0.72			0.70
<i>other</i>	0.74				0.63
<i>debate</i>					0.72
<i>full-debate</i>					0.74
<i>wiki</i>	0.83	0.73	0.79	0.70	
RoBERTa	<i>conseq</i>	<i>other</i>	<i>debate</i>	<i>full-debate</i>	<i>wiki</i>
<i>conseq</i>		0.72			0.73
<i>other</i>	0.46				0.50
<i>debate</i>					0.83
<i>full-debate</i>					0.88
<i>wiki</i>	0.89	0.82	0.86	0.78	
DeBERTa	<i>conseq</i>	<i>other</i>	<i>debate</i>	<i>full-debate</i>	<i>wiki</i>
<i>conseq</i>		0.85			0.82
<i>other</i>	0.89				0.80
<i>debate</i>					0.85
<i>full-debate</i>					0.90
<i>wiki</i>	0.91	0.86	0.89	0.80	
StArCon	<i>conseq</i>	<i>other</i>	<i>debate</i>	<i>full-debate</i>	<i>wiki</i>
	0.73	0.64	0.69	0.62	0.64

Table 4.9: Cross-domain evaluation for stance detection: The training sets are presented in the first column, the test sets in the head row. The scores are accuracy values.

when trained on *conseq*. Further, the more training data there is provided, the better the results are for *wiki* – even if the training data is noisier (*full-debate*). Lastly, in terms of stance detection, StArCon is considerably worse than BERT-like models if they are provided enough training data.

4.5.3 Error Analysis

In order to identify the systematic mistakes done by the transformer based models, we manually looked at some instances which got classified wrongly by DeBERTa. We could not find any obvious similarities among the misclassified instances, except that about one third of the misclassified instances are either very hard or even impossible to solve. Further, about one quarter of the misclassified instances involve factual background knowledge which is needed to properly solve them, e.g.,

- Instant replay in baseball – Fallible umpire calls are part of the drama of baseball
- DREAM Act – Republican opposition alienates Latino vote.

Quantitatively, DeBERTa and StArCon tend to misclassify the same instances. The tendency is consistent among the different datasets: The number of commonly misclassified instances is 1.4 to 1.8 as high as one would expect by random based on the systems’ accuracy scores. This by itself is quite natural as the datasets contain instances that are very hard to solve. But furthermore, this tendency is a bit stronger if we only consider the cases where StArCon did find an explanation for its result, i.e., if it did not need to use its sentiment-based fall-back strategy. Then, the quotient lies between 1.5 and 2.5. Thus, possibly, if the coverage of instances where we can apply similar meaningful rules would be higher, the predictions would be more consistent with those of DeBERTa.

Lastly, we manually analyzed instances which either both DeBERTa and BERT solved successfully but StArCon failed to, or vice versa. We did not find any mentionable patterns or biases.

4.6 Conclusion

In this chapter, we propose a method to extract effect triples from the effect premise of an argument from consequences by exploiting grammatical dependencies and lexicons to identify effect words and their impact. With the effect triple, we can concretely identify ACTION and CONSEQUENCE and fully reconstruct the argumentation scheme by inferring CONSEQUENCE’s alignment (judgment premise) and thus the conclusion. We consider this to be an important step towards enthymeme reconstruction, though there are two major limitations: First, in our experiments, we identify consequences only for 60% of the arguments from consequences. Second, we do not address arguments where the effect premise is implicit. However, we hypothesize that at least if the judgment premise and the conclusion are explicit, inferring the effect premise can be done rather straightforwardly by identifying CONSEQUENCE in the judgment premise and ACTION in the conclusion. The major challenge might be to detect such arguments as arguments from consequences, because the commonly used trigger, the effect relation, is implicit. Our effect graph which we introduce in the next chapter might help to overcome this challenge.

Further, with StArCon, we propose a stance detection method based on our effect triple extraction. StArCon is fully unsupervised and provides explanations for its classification results in the form of the effect triples. For our evaluation, we annotated arguments from *Debatepedia* regarding their stance and whether they involve consequences

or not. While StArCon’s performance on the stance detection task is comparable to BERT in some settings, it is inferior to more recent transformer based models.

Even aside their use for enthymeme reconstruction and stance detection, we consider the extracted effect triples to be valuable. Particularly, they contain effect relations. Such effect relations have been used as external knowledge by Al Khatib et al. (2021) and Yuan et al. (2021). They are also the foundation of our explanations for the effect premise in the following chapter.

Chapter 5

Explaining the Effect Premise: The Effect Graph

¹ In this chapter, we go beyond reconstructing an argument from consequences and propose a method to identify possibly explanations for the effect premise. In short, the aim is to explain why ACTION causes CONSEQUENCE. To motivate our approach of explaining the effect premise, we look at the following example:

Example 13. *Legal abortions protect women.*

First, we note that it is not possible to find the one and only explanation for why one would claim that legal abortions protect women. Instead, there exist many different possible explanations and, from merely reading the premise, we cannot know which of these the statements the author had in mind. Some examples are listed in table 5.1. Thus, our goal is not to reconstruct the original explanation, but to find meaningful ones.

For automatically generating possible explanations, we choose an approach that is specific for explaining relations of the type $A \overset{+}{\rightarrow} B$ or $A \overset{-}{\rightarrow} B$. $A \overset{-}{\rightarrow} B$ means that entity A expresses a negative effect on entity B . We follow Al-Khatib et al. (2020) by calling these relations *effect relation*. They can be seen to be a subset of the effect triples introduced in the previous chapter. The effect relation for example 13 is *legal abortions* $\overset{+}{\rightarrow}$ *women*. In terms of the argumentation scheme, instance A of the effect relation

¹This chapter is adapted from Kobbe, J., Hulpus, I., and Stuckenschmidt, H. (2023). Effect graph: Effect relation extraction for explanation generation. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 116–127, Toronto, Canada. Association for Computational Linguistics.

The corresponding resources are publicly available at <https://github.com/dwslab/Effect-Graph>.

Explanations	
1	Abortions protect women from the harm caused by giving birth and being pregnant.
2	Abortions prevent long term damage caused by complications during the pregnancy and birth process.
3	Legal Abortions strengthen the women’s right to self-determination.
4	Abortions release women from the financial burden of raising a child.
5	Abortions can protect girls from becoming mothers too early.

Table 5.1: Some possible explanations for example 13.

Effect-effect-explanations	
1	Abortions $\bar{\rightarrow}$ harm $\bar{\rightarrow}$ women
2	Abortions $\bar{\rightarrow}$ long term damage $\bar{\rightarrow}$ women
3	Legal Abortions $\overset{+}{\rightarrow}$ right to self-determination $\overset{+}{\rightarrow}$ women
4	Abortions $\bar{\rightarrow}$ financial burden (of raising a child) $\bar{\rightarrow}$ women

Table 5.2: Formalized effect-effect-explanations for example 13.

matches ACTION, while CONSEQUENCE consists of both B and the effect expressed upon it ($\overset{+}{\rightarrow}$ women or, more concretely, *protect women*).

Our core idea to explain the effect premise is to add more detail to the relation by finding an instance C such that $A \rightarrow C \rightarrow B$. Because of the structure of such an explanation, we call it *effect-effect-explanation*. Of course, with such a formalized explanation, we cannot capture all the details in the explanations in table 5.1. But we can capture some key aspects of these explanations and describe the explanations in a well-defined way that easily allows for further processing in downstream tasks. Table 5.2 shows possible formalized versions of explanations 1 to 4.

Effect-effect-explanations are, however, still very limited in their nature. While we cannot fully overcome this limitation, we show that it is possible to expand upon them for instance by incorporating lexical knowledge: Given $A \rightarrow B$, an explanation could also be $(A \rightarrow C, C \text{ instanceOf / hypernym / synonym } B)$ or, vice versa, $(A \text{ instanceOf / hypernym / synonym } C, C \rightarrow B)$. Analogously, we call these *effect-lexical-explanation*. An example is given in table 5.3.

The main challenge for both of the explanation schemes, effect-effect-explanation and effect-lexical-explanation, is to get the additional information (i.e., C and its links to A and B). For the lexical relations, we again use WordNet. For the effect relations, the only potentially appropriate resource we are aware of is the KG presented in Al-Khatib et al. (2020) and Al Khatib et al. (2021). We create our own KG containing effect rela-

Effect-lexical-explanation		
5	Abortions $\xrightarrow{+}$ girls $\xrightarrow{\text{hypernym}}$ women	

Table 5.3: Formalized effect-lexical-explanations for example 13.

tions, which we call *effect graph*, and improve upon the existing KG by (i) using a new precision-focused unsupervised effect relation extraction method; (ii) including effect relations not only from argumentative texts, but also from encyclopedic ones (iii) having substantially more nodes and edges; (iv) making the effect graph publicly available.

For the effect relation extraction, we propose an adopted version of StArCon to extract effect relations from both argumentative and encyclopedic texts. Other than in the previous chapter, we consider precision to be more important than recall for the information extraction, most importantly because we expect this to benefit the quality of the explanations, but also because a low recall can be compensated to a certain degree by applying the extraction method on more data.

In the following, we describe the generation of the effect graph in section 5.1. Then, we propose a method to generate and rank explanations for effect relations in section 5.2. We evaluate both the effect graph and the explanation generation in section 5.3. Lastly, we conclude with a discussion in section 5.4.

5.1 Effect Graph Generation

Our aim is to generate a graph where the nodes are concepts such as *global warming*, *CO2 emissions*, *solar panel*. The edges represent the effect relations and indicate either a negative or positive effect from the source to the target node, e.g., (*solar panel*) $\xrightarrow{+}$ (*CO2 emissions*). We also store the concrete word indicating the effect. In the previous example, this could be for instance *reduce* or *prevent*. In order to extract effect relations from text, we propose a method which is based on StArCon. We refer to it as Effect Relation Extractor (EREx). The main differences between StArCon and EREx, besides EREx not predicting a stance, are due to the following reasons: First, StArCon requires a topic specifically to identify an effect triple’s subject, while EREx should be applicable without specifying a topic. Second, in order to predict a stance for as many inputs as possible, StArCon also exploits less reliable patterns. EREx, however, is not expected to find an effect relation in every statement it is applied on and thus relies only on the more robust patterns to extract effect relations. Third, StArCon requires the effect triple’s object to have a sentiment in order to calculate the stance, which is not

Pattern	Interpretation	Example
1 $P \xrightarrow{*} O$	O object of P	Insurance mandates violate the rights of employers.
3 $P \xrightarrow{\diamond} S$	S subject of P	Holocaust denial is inherently discriminatory and damaging.
5 $NegP \xrightarrow{pobj} X$	X is negated	Free speech without Fairness Doctrine can harm policy-making
6 $X \rightarrow NegP \wedge \nexists NegP \xrightarrow{pobj}$	X is negated	W/o more troops, Afgh will become terrorist haven
7 $X \xrightarrow{neg}$	X is negated	Solar energy does not damage air quality.

$* \in \{dobj, cobj, nsubjpass, csubjpass\}; \diamond \in \{nsubj, csubj\};$
 $NegP$ stands for *negative preposition*

Table 5.4: Dependency graph patterns for EREx.

needed for EREx. Because of that and reason one, the subjects and objects derived by the patterns are no longer controlled for by either linking to the topic or a sentiment lexicon, so we pose other restrictions on both of them.

5.1.1 Effect Relation Extraction (EREx)

We use a subset of the dependency parse patterns used by StArCon. The patterns are presented in table 5.4. Using these patterns, we look for triples (S, P, O) such that the predicate P has subject S and object O and where none of S , P and O are negated. In order for the triple to qualify as effect relation, P has to be an effect word which we identify by applying ECF (see chapter 4.1.1) with a threshold of ± 0.2 . The effect relation's subject, which we denote by A , is then the statement's substring which is represented by the dependency parse's subtree whose root is S . Analogously, the object B is the statement's substring represented by the subtree whose root is O . Thereby, leading articles are ignored. A and B , which will become the effect graph's nodes, are required to be non-stopword NPs which link to an entry in Wikipedia. This ensures that they are meaningful entities in different contexts. If all these requirements are met, we consider $A \xrightarrow{P} B$ to be an effect relation.

5.1.2 Graph Construction

For building the effect graph, we extract effect relations from the following three text resources:

Debatepedia We already used Debatepedia for evaluating StArCon where we annotated a part of the sentences with whether they relate to consequences or not. This is valuable as we expect arguments from consequences to contain effect relations more often than other arguments. For the purpose of relation extraction, we do not restrict ourselves to the arguments' claims, but instead parse all the arguments from the featured debates.

Debate.org As Debatepedia is rather small, we also use Debate.org (Durmus and Cardie, 2018, 2019) to extract effect relations from a large argumentative text basis. In Debate.org, two persons engage in a debate about a certain topic and present their arguments and counter arguments over three rounds. Presumably, most users engage in debates for fun or for training their argumentation skills. Thus, Debate.org contains debates about socially relevant topics, but also silly topics like *Are Hot Dogs Sandwiches?* or *Is Joe Biden a zombie*. Also, the arguments vary a lot in quality as well as seriousness and are sometimes factually wrong. However, because of the low threshold of starting and participating in a debate, Debate.org covers a large amount of topics and contains many arguments.

Simple Wikipedia Lastly, we use an encyclopedic text resource to also capture non-argumentative knowledge which can be relevant for explaining arguments. To save computational resources and increase the accuracy of the extraction process, we use the Wikipedia version in simple English.

Both argumentative text resources mainly contain defeasible arguments. Thus, the effect relations which we extract from them and, consequentially, the effect graph should not be treated as facts.

After extracting the effect relations from text, we remove duplicates. We only consider an effect relation to be a duplicate, if it was extracted from the same sentence in the same resources twice, which most often happens because of citations. We intentionally keep effect relations that are identical except for the sentence they were extracted from because this might indicate that the effect relation is especially relevant.

For building the effect graph, we connect the extracted effect relations as follows: The lemmas of the subjects A and the objects B become nodes. We add one edge between A and B for every respective effect relation we extracted. Since we do not collapse the edges to not lose any information, the resulting graph is expected to contain multi-edges. Figure 5.1 gives a brief impression of how the graph looks like.

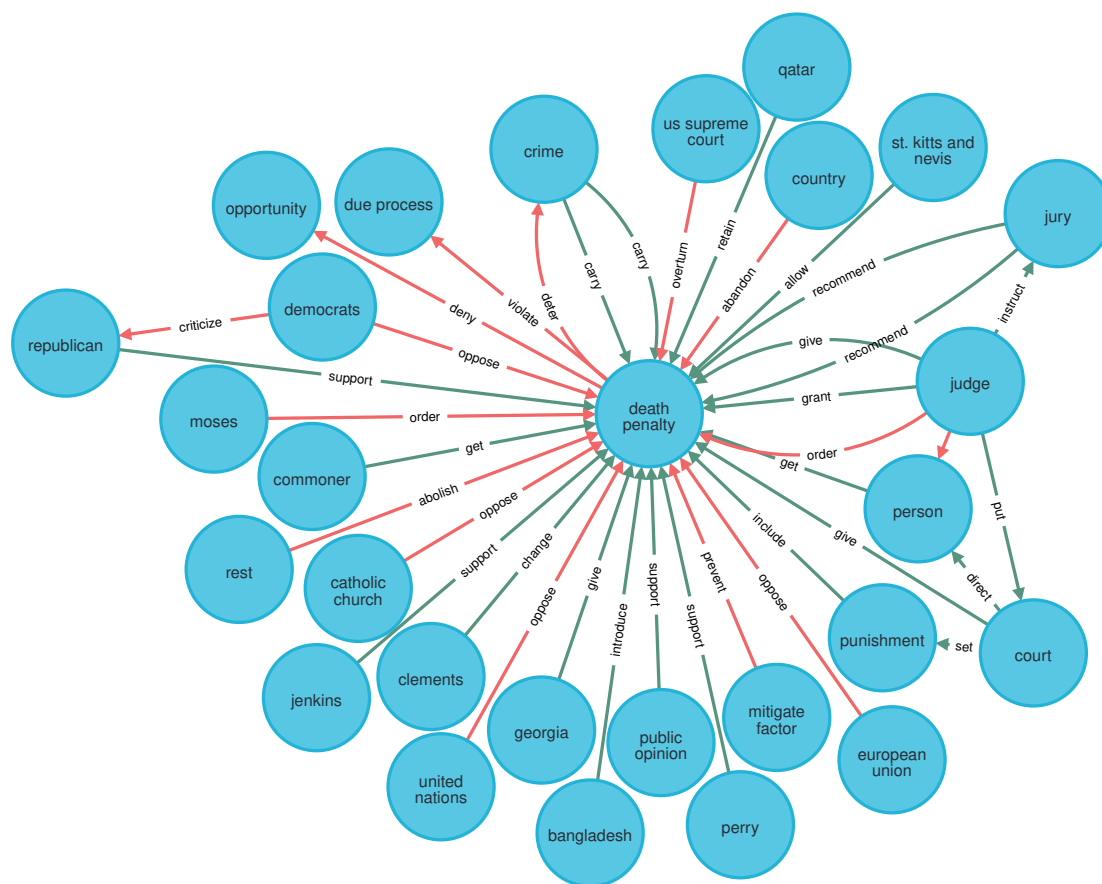


Figure 5.1: Effect graph snapshot: Nodes neighboring *death penalty* and the relations between them, extracted from Debatepedia and Simple Wikipedia only.

5.2 Explanation Generation

For generating explanations, we use the lemmatized effect graph generated by EREx. As outlined in the introductory section, we envision two different types of explanations which we will describe separately in the sections 5.2.1 and 5.2.2. Afterward, we introduce a measure to rank the potential explanations in section 5.2.3.

5.2.1 Effect-Effect-Explanation

For an effect-effect-explanation to be a meaningful explanation, the polarities have to fit the relation we aim to explain. We consider the latter to be a necessary requirement, but it certainly is not sufficient as we will see in the evaluation. Table 5.5 shows the specific combinations of polarities which make sense in this context.

To generate explanation candidates as indicated in table 5.5, we use the effect graph

To explain	Explanation candidate
$A \xrightarrow{+} B$	$A \xrightarrow{+} C \xrightarrow{+} B$
	$A \xrightarrow{-} C \xrightarrow{-} B$
$A \xrightarrow{-} B$	$A \xrightarrow{+} C \xrightarrow{-} B$
	$A \xrightarrow{-} C \xrightarrow{+} B$

Table 5.5: Effect-effect-explanation: Explanation candidates for positive and negative effect relations.

in a straight forward way by querying for paths of length two between the instances of interest with appropriate edge polarities. As a result, we get a list of explanation candidates.

For our abortion example from the introduction (see example 13 and table 5.2), this list includes 370 explanation candidates, though many of them are similar to each other because of our loose definition of duplicates. Instead of listing all candidates, we list all the interim nodes C used within the explanation candidates: *, choice, country, fetus, god, man, nothing, order, people, person, pregnancy, right, sex, society, t, unwanted pregnancy, woman 's rights. While some of the concepts mentioned are useful for explaining why abortions protect women, others are not. To distinguish between meaningful concepts and noise, we introduce a node ranking in section 5.2.3.

5.2.2 Effect-Lexical-Explanation

Sometimes, we need additional lexical knowledge for explaining an effect triple. As mentioned previously, we use WordNet to incorporate some of the potentially relevant lexical knowledge. Concretely, this includes hyperonymy, meronymy and synonymity.² While synonymity is strictly bidirectional, the hyperonymy and meronymy relations have different names depending on the direction (hypernym vs hyponym, meronym vs holonym).

To extract explanation candidates for $A \xrightarrow{\pm} B$, we again look for instances C , considering the following cases: $A \xrightarrow{\pm} C \xrightarrow{\text{WN}} B$ and $A \xrightarrow{\text{WN}} C \xrightarrow{\pm} B$. The polarities have to be identical and $\xrightarrow{\text{WN}}$ indicates one of the lexical relations mentioned above.

For our abortion example from the introduction (see table 5.3), we find 10 different explanation candidates. Half of them argue that abortions are good for mothers in some way, and mother is a hyponym for woman. While being trivial, we still think that there

²Hyperonymy describes *[is a]* relations, like *bike [is a] vehicle*. Meronymy describes *[part of]* relations, like *wheel [part of] bike*.

is a benefit in this explanation. It states correctly that the positive effect of the abortion is on the mother (and not on the fetus, for instance) and finds the relation between *mother* and *woman*. The other five explanation candidates use the interim nodes *people*, *action*, *failure*, *man* and none of these explanation candidates seem helpful, which further motivates our node ranking in section 5.2.3.

Oftentimes, when working with WordNet, disambiguation of the different word senses is an issue. However, we do not consider it a major problem in our context: We already know that our instances A and B are in a relation to each other. By looking for short paths like $A \xrightarrow{+} C \xrightarrow{\text{hyponym}} B$, where we know that A and B are related to each other, we assume that in most cases where there is a relation between any of the synsets belonging to C and B , it is between those having the correct meaning. Thus, the context we provide is supposedly self-disambiguating for the majority of cases.

5.2.3 Explanation Candidate Filtering

Since the proposed methods to generate explanations may result in a list of explanation candidates of varying quality, we further propose a simple means of ranking them which is inspired by tf-idf. The idea is to measure the importance of the interim node C based on its degree in the effect graph and in the queried subgraph. The node's degree should be low in the effect graph, as this indicates specificity. As queried subgraph, we denote the graph containing all paths of length 2 between A and B . At least one edge per path has to be from the effect graph, while the other one can potentially be a relation from WordNet. Thus, the graph includes all explanation candidates. Since the original effect graph has many multi-edges, each edge representing one argument, the subgraph also can have multi-edges. In this case, the more edges between two instances, i.e., the higher C 's degree, the more often the relation between these two instances was mentioned. This indicates both a higher confidence that the relation was extracted correctly and a higher relevance. The core idea for measuring relevance is the quotient of these two quantities, deg_s denoting the degree in the subgraph and deg_e the degree in the effect graph:

$$\frac{deg_s(C)}{deg_e(C)} \quad (5.1)$$

This quotient, however, does not respect the absolute quantities and will thus lead to the same score for C having degree 1 in both graphs and having degree 5 in both graphs. However, we consider the latter to be considerably better. In order to account for that,

we apply the idea of additive smoothing and increment the denominator by 1:

$$\frac{deg_s(C)}{deg_e(C) + 1} \quad (5.2)$$

Lastly, we distinguish between effect-effect- and effect-lexical-explanations: In the first case, we consider it to be better to have a medium in- and out-degree rather than a high in- and low out-degree or vice versa, because both the connection $A \rightarrow C$ and $C \rightarrow B$ have to be sufficiently specific and relevant. In other words, the explanation is only as good as its weakest part. Consequently, we define C 's importance as follows:

$$importance(C) = \frac{indeg_s(C)}{indeg_e(C) + 1} \cdot \frac{outdeg_s(C)}{outdeg_e(C) + 1} \quad (5.3)$$

Considering effect-lexical-explanations, we are only interested in either C 's out- or in-degree. For better comparability, we use the square of the relevant quotient to measure the importance. However, one might also want to give more or less weight to this explanation type overall, or to give more or less weight to specific relations in WordNet.

When applying the importance measure on the abortions example, the five most important nodes are in descending order: unwanted pregnancy, woman's rights, mother, fetus, pregnancy. The corresponding explanation via *unwanted pregnancy* unfortunately does not make sense due to an extraction mistake, although the concept seems to be ranked that high for good reason. The other explanations are valid though. We already discussed the one via *mother* in section 5.2.2. The others argue that (i) abortions kill fetuses which in turn harm, damage or endanger the woman; (ii) abortions end pregnancies which also harms the woman; (iii) abortions support women's rights which in turn are good for women.

In the following, we evaluate the effect relation extraction, the effect graph itself and the explanation generation.

5.3 Evaluation

We evaluate the effect graph and its use for explanation generation as follows: In section 5.3.1, we evaluate the effect relation extraction process using the subtasks defined by Al-Khatib et al. (2020). In section 5.3.2, we evaluate the extracted effect graph itself in terms of precision and recall. Lastly, we evaluate the explanation generation in

section 5.3.3.

5.3.1 Effect Relation Extraction Evaluation

Al-Khatib et al. (2020) propose several subtasks for effect relation extraction. These subtasks include:

- **Relation Classification:** Classify whether a statement does contain an effect relation.
- **Relation Type Classification:** Classify whether the effect relation is positive or negative.
- **Subject Identification:** Identify the effect relation’s subject.
- **Object Identification:** Identify the effect relation’s object.

Data For the evaluation, we use the dataset published by Al-Khatib et al. (2020) which contains crowd annotations for the individual subtasks. The dataset contains 4740 statements of which 1736 are annotated to contain an effect relation. 74% of the effect relations are positive and 23% are negative. Since the IAA is only moderate, the results should be interpreted carefully. Especially the annotation of the effect relations’ objects is affected by inconsistencies and ambiguities such as articles sometimes being part of the object and sometimes not.

Baseline As a baseline, we use the results reported in Al-Khatib et al. (2020). For the subject and object identification, Al-Khatib et al. rely on the OpenIE approach of Stanovsky et al. (2018) which outperformed an alternative based on the semantic role labeling approach of He et al. (2017). For the two classification tasks, they train a support vector machine using lexical, syntactical, sentiment-based and semantic features. In the latter category, the connotation frame lexicon, which our ECF is based on, is included.

To make the comparison fair, we slightly adopt EREx such that it predicts a relation type and identifies concepts even if it does not detect an effect relation. The results are presented in table 5.6.³

Concerning relation classification, EREx misses effect relations considerably more often than it wrongly predicts one (1582 vs 174 instances), which fits our focus on precision rather than recall. When counting only such instances where EREx extracts

³Since the train-test-split used by Al-Khatib et al. (2020) is unknown to us, we use the full dataset for the evaluation. Thus, unfortunately, the results are not directly comparable.

Subtask	Measure	Al-Khatib	EREx
Relation Classification	macro F1	0.79	0.65
Relation Type Classification	macro F1	0.77	0.77
Subject Identification	accuracy	0.69	0.71
Object Identification	accuracy	0.28	0.35

Table 5.6: Effect relation extraction evaluation.

a relation, it correctly detects its polarity in 85%, the subject in 80% and the object in 41% of the instances. While both models’ scores of identifying the object are low, this can be explained at least partly by the measure: The object is considered to be wrong if it is off by one word, even if it is an article.

5.3.2 Effect Graph Evaluation

Our evaluation of the effect graph itself consists of three parts. First, we present graph statistics. Afterwards, we evaluate both precision and recall. In this context, precision expresses the chance that a randomly selected edge of the graph is correct. We consider a statement to be correct if it is in accordance with the statement it was extracted from. Recall on the other hand measures the chance that a given effect relation is contained in the graph.

As baselines, we build the effect graph as described in section 5.1.2, but using different extraction methods. We use the **OpenIE** implementation which is part of Stanford CoreNLP (Manning et al., 2014; Angeli et al., 2015) to extract subject-verb-object triples, applying a confidence threshold of 0.9. We accept such triples as effect relations where the verb is an effect word and the subject and object link to Wikipedia pages.

Further, we use a version of EREx where we do not require the subject and object to link to Wikipedia, denoted by **EREx***. We expect this version to have a higher recall, but also more noise.

For the comparison of graph statistics, we compare our graph to the numbers reported in Al Khatib et al. (2021). We consider both (i) their graph based on the annotated data of Al-Khatib et al. (2020) and (ii) the graph they generate by applying the extraction method presented in Al-Khatib et al. (2020) on *args.me* and *kialo.com*⁴. We collapse these graphs to one single graph denoted by **AKG** (Argumentation Knowledge Graph, or Al-Khatib’s Graph).

⁴<https://args.me> and <https://kialo.com>

Dataset	Size	Number of effect relations			
		EREx	EREx*	OpenIE	AKG
Debatepedia	5 MB	1,653	8,833	9,931	9,100
Debate.org	728 MB	150,359	669,900	1,173,879	
Simple Wikipedia	190 MB	43,676	193,916	290,352	
args.me+kialo.com					14,643

Table 5.7: Effect relation extraction statistics.

	EREx	EREx*	OpenIE	AKG
# Nodes	53,773	734,905	129,534	~ 23,000
# Edges	195,688	872,649	1,474,499	23,743
# Positive edges	157,161	729,974	1,250,965	17,907
# Negative edges	38,527	142,675	223,534	5,836
# Connected node pairs	126,238	733,632	603,054	

Table 5.8: Effect graph: Statistics.

Effect Graph Statistics

Table 5.7 shows the number of edges, i.e., extracted effect relations per dataset. Note, however, that AKG is based on a different subset of Debatepedia than the others. Table 5.8 contains some basic statistics of the effect graph. The number of connected node pairs is included because of the high ratio of multi-edges. We consider (A,B) and (B,A) as the same node pair. Table 5.9 shows the number of overlapping nodes between the different effect graph versions.

Overall, when comparing our three extraction methods, using OpenIE results in the largest graph and using EREx in the smallest, though the graph is still considerably larger than AKG. The fact that OpenIE extracts fewer nodes than EREx* is likely due to the required linking to Wikipedia. The linking is probably also the reason for why the graphs generated by EREx and OpenIE are considerably denser than the one generated by EREx* and AKG.⁵ For all four graphs, there are considerably more positive than negative effect relations.

Effect Graph Precision

We evaluate the effect graph’s precision a posteriori. For this purpose, we randomly select 250 edges per graph. For each, we annotate whether it was extracted correctly,

⁵Presumably, AKG contains considerably less multi-edges. However, the conclusion still holds when comparing the number of connected node pairs in our graphs to the number of edges in AKG.

	EREx	EREx*	OpenIE
EREx	–	52,821	43,527
EREx*	52,821	–	63,827
OpenIE	43,527	63,827	–

Table 5.9: Effect graph: Node overlap.

given the original statement (*yes, rather yes, unsure, rather no, no*). We do both an expert annotation by the author and crowd annotations via mturk.

Instructions We require the crowd workers to successfully pass an instruction before working on the task. The instruction consists of a short description of the task, two examples with comments, three instances which had to be annotated correctly, and an optional field where the workers could write comments. The details are presented in figure 5.2.

Overall, the task should be as intuitive as possible. For this purpose, we did not show the concrete verb of the effect relation, but just the effect’s polarity. Instead of explaining that we are not interested in modality, we framed the polarity as “(may) negatively affect”. Also, we decided against colorcoding the polarities because this might increase the risk of confusion with sentiment. Indeed, this potential confusion is one of the main reasons for having the instructions. We addressed it in example 1 and instances 2 and 3 which are included in the description of figure 5.2. In example 1, most would likely agree that *ending war* is desirable, thus we highlight that the effect which is expressed on *war* is a negative one. Instance 2 also contains a negative effect on something bad (*coal power* reducing *CO2-emissions*) while instance 3 addresses a positive effect on something bad (*current EU policy* leading to a *financial crisis*). The other two cases, positive and negative effects on something good, are unproblematic. Example 2 and instance 1 are for emphasizing that not only the effect matters, but that the subject and object also have to be correct.


Annotation Process Similarly to the crowd annotation presented in section 4.3, we only accept workers who live in the US and have a HIT approval rate greater than 98% and more than 10,000 approved HITs in total. Additionally, they have to have passed the instructions with three correct answers out of three. As the cases in the instructions were not ambiguous, we count *rather yes* and *rather no* as wrong answers, as well as *unsure*. Overall, 9 out of 50 workers passed the instructions. While this is fewer than we expected, it also confirms our decision to filter out workers who do not understand

Each HIT, you will be presented a **Statement** from which a **Relation** was extracted automatically. The Relation is expected to capture some sort of positive or negative effect between two of the statement's instances.

Your task is to judge whether the extraction was successful. Successful means that the Relation can be considered to be correct when assuming that the Statement itself is correct.

Example 1

Statement: Scientists found out that unicorns can end any war.

Relation: unicorns  war

Obviously, the statement is made up. But for this task, we assume it to be true.

Consequently, the Relation is **correct**: The effect which is expressed on *war* is a negative one (it may be *ended* by unicorns).

Other words that trigger negative effects are for instance *decrease, damage, forbid, ban, reduce,* Positive effects are triggered by words such as *increase, help, permit, cause, create,*

Example 2

Statement: Scientists found out that unicorns can end any war.

Relation: scientists  war

This time, the Relation is **not correct**: It is not the *scientists* who have a negative effect on *war*, but the *unicorns*.

Your turn!

Statement: Throughout history, nuclear weapons have killed many innocents.

Relation: history  innocents

Assuming the Statement is correct, is the Relation also correct?

- ☐ No
- ☐ Rather no
- ☐ I am unsure
- ☐ Rather yes
- ☐ Yes

Figure 5.2: Instructions for the crowd workers. Analogously to the last instance, there are two additional ones: (2) *Using more coal power would reduce our CO2-emissions by a large amount.* [*coal power* $\xrightarrow{+}$ *CO2-emissions*]; (3) *The current EU policy will lead to a financial crisis.* [*current EU policy* $\xrightarrow{+}$ *financial crisis*].

the task or are unreliable in general.

We have a total of 750 instances to be annotated. Each instance is annotated by three crowd workers and one expert. Overall, seven of the nine qualified workers did actually address the task. Of these seven workers, three did annotate the vast majority of the instances (747, 739 and 650 respectively).

The instances were presented one by one, with the instructions still being available. Again, the workers had the chance to write comments. This opportunity was used only two times, once for stating that the relation does not make sense and once for stating that the statement itself is confusing since it misses the proper context.

Agreement To get a first impression, we counted that the all three crowd annotators agreed in about half of the cases when treating *rather yes* as *yes* and *rather no* as *no*. This indicates at least some agreement as by chance, we would expect them to agree in at most one quarter of the cases.⁶

For measuring IAA properly, we consider different options for (i) how to treat the five labels, (ii) whether to consider the actual label distribution, and (iii) how to handle the expert as opposed to the crowd workers.

Concerning the five labels, we treat them either as *polarities*, mapping *rather yes* to *yes* and *rather no* to *no* for calculating categorial agreement. Or, alternatively, we map them to *scalars* as indicated in table 5.10 for calculating scalar or rank agreement.

categorial label	value
yes	2
rather yes	1
unsure	0
rather no	-1
no	-2

Table 5.10: Mapping categorial answers to values.

The mapping allows us to intuitively combine multiple labels by computing their mean which is relevant for generating the final label to ultimately measuring the precision. But it also enables us to measure the agreement of the combined label with the expert annotator (*mean+expert*). Alternatively, we include the expert annotator as an equal to the crowd annotators (*crowd+expert*) or exclude him for objectivity sake

⁶To get an upper limit of what to expect by chance, we assume that *unsure* is way less popular as an answer than the other two. If it would be an equally popular choice, then the expected agreement by chance would be even lower.

		crowd	crowd+expert	mean+expert
polarities	Fleiss	0.15	0.20	0.26
	Randolph	0.47	0.44	0.44
scalar	Krippendorff	0.20	0.26	0.34
	Pearson			0.57
	Spearman			0.56

Table 5.11: Agreement scores for effect relation evaluation.

(*crowd*). For mapping back from numbers to labels, we always round up positive values and round down negative values. This way, the labels *yes* and *no* are only provided if there are no opposing polarities and the label *unsure* is given as rarely as possible.

We calculate the following agreement scores:

- **Fleiss Kappa** for categorial agreement respecting the label distribution.
- **Randolph Kappa** (Randolph, 2005) for categorial agreement without respecting the label distribution.
- **Krippendorff Alpha** (Krippendorff, 2011) for scalar agreement, especially in the *crowd* and *crowd+expert* setup as it allows for multiple annotators.
- **Pearson Correlation** for scalar agreement in the *mean+expert* setup, using the mean as is.
- **Spearman Correlation** for rank agreement in the *mean+expert* setup, mapping the mean to labels.

The scores are presented in table 5.11. Overall, the agreement is rather weak. In the polarities setup, we note two things: First, there is a big difference between Fleiss and Randolph which we explain by the fact that the crowd workers tended to annotate *yes* or *rather yes* way more often than *no* or *rather no* (552 vs 173 according to *mean*). Second, for Fleiss, the more the expert is involved, the higher the scores, while for Randolph it is rather vice versa. This tendency might be explained by the fact that the expert annotated *yes* or *rather yes* in only 341 cases, which is even less often than *no* or *rather no*. So the expert reduces the imbalance between these two labels which in turn causes Fleiss and Randolph to approach each other.

For the scalar agreement, the scores seem to be a bit better which makes sense as only in this scenario the rank of the labels is considered properly. As the agreement scores also seem to be better when including the expert, despite his tendency to annotate *yes* and *rather yes* considerable less often than the crowd annotators, this might indicate

	exclusive		inclusive	
	total	precision	total	precision
OpenIE	115	0.83	237	0.70
EREx	132	0.98	246	0.80
EREx*	130	0.95	242	0.79

Table 5.12: Effect graph precision, based on crowd annotations.

	exclusive		inclusive	
	total	precision	total	precision
OpenIE	186	0.38	241	0.34
EREx	174	0.54	243	0.54
EREx*	175	0.48	248	0.46

Table 5.13: Effect graph precision, based on expert annotations.

that at least one of the crowd workers was rather unreliable. However, we still conclude that the IAA is moderate at best which we have to consider when interpreting the results.

Results The precision scores are calculated by dividing the number of correctly extracted effect relations by the sum of the numbers of correctly and incorrectly extracted ones. As for what we consider a correctly extracted effect relation, we again consider different settings to provide a full picture. For one, we use either the expert label or the aggregated crowd label. Further, we either consider only the labels we are confident about, *yes* and *no* (denoted by *exclusive*), or we again aggregate *yes* and *rather yes* as well as *no* and *rather no* (denoted by *inclusive*). We never consider the relatively few cases where the (aggregated) label is *unsure*. The results are shown in table 5.12 and table 5.13.

The expert’s tendency to annotate *yes* considerably less often than the crowd workers is reflected by the overall lower precision scores. Despite this large difference of the scores, the tendency among the datasets is consistent for the crowd workers’ and the expert’s annotations: EREx and EREx* clearly outperform OpenIE, while EREx seems to be at least slightly better than EREx*. This was to be expected as EREx is more restrictive in selecting subjects and objects than EREx*.

We conclude that EREx and EREx* are most likely more precise than the OpenIE baseline, but whether or not they are precise enough for our envisioned use case is yet to be shown.

Ex. 1	Calorie counts eliminate ability of restaurants to be spontaneous.
a	(Calorie counts) [-eliminate] (ability of restaurants to be spontaneous)
b	\supset (Calorie counts) [-eliminate] (ability of restaurants)
c	\supset (Calorie counts) [-] (restaurants)
Ex. 2	Circumcision creates risk of infections in infants
a	(Circumcision) [+creates] (risk of infections)
b	\equiv (Circumcision) [+creates] (infections)
Ex. 3	Assassinations protect publics from terrorism; even while it's hard to measure
a	(Assassinations) [+protect] (publics)
b	$\not\equiv$ (Assassinations) [-protect from] (terrorism)
Ex. 4	Network neutrality damages competition and niche suppliers
a	(Network neutrality) [-damages] (competition and niche suppliers)
b	\equiv [(Network neutrality) [-damages] (competition)]
c	$\not\equiv$ [(Network neutrality) [-damages] (niche suppliers)]

Table 5.14: Examples: Effect relation annotation for recall evaluation.

Effect Graph Recall

For evaluating recall, we check whether the graph does contain effect relations which we would expect it to contain. In order to do so, we build an evaluation dataset: We choose one random argumentative claim per topic from our Debatepedia dataset containing only arguments related to consequences, as we consider them to contain effect relations more often than other arguments. This results in 180 claims. From each claim, we manually extract all effect relations which we consider reasonable. If there is more than one possible effect relation for a claim, we annotate whether they are either equivalent to (\equiv), disjoint to ($\not\equiv$), or part of (\supset) the other ones. Table 5.14 shows some examples which we will briefly discuss.

In example 1, there exist three effect relations which make sense to extract and which differ only in the concreteness of the object, effect relation *a* being the most concrete and effect relation *c* the least. Note that the effect verb *eliminate* is only correct when mentioning the *ability* of restaurants. Still, the statement implicitly also expresses that *calorie counts* negatively effect restaurants, which is why in effect relation *c*, there is no effect verb annotated.

Example 2 briefly shows a case where there exist two effect relations which are roughly equivalent in terms of the information they contain. In contrast, in example 3 exist two completely distinct effect relations, though the second one is rather implicit.

Example 4 is a bit more complex: effect relation *a* is very concrete, but you can also

	total		per statement	
	full graph	w/o Debatepedia	full graph	w/o Debatepedia
OpenIE	0.07	0.04	0.14	0.09
EREx	0.05	0.03	0.09	0.06
EREx*	0.14	0.03	0.28	0.06

Table 5.15: Effect graph recall.

split this effect relation into two effect relations which are distinct from each other but equivalent to effect relation a when considered together.

For calculating recall, we use two straightforward formulas: We either divide the number of the ground truth effect relations which are contained in the effect graph by the total number of ground truth effect relations (*total*), or we divide the number of claims for which at least one ground truth effect relation is contained in the effect graph by the number of claims in the dataset (*per statement*). Further, we optionally exclude the effect relations which were extracted from Debatepedia from the effect graph (*w/o DP*). Though it is unclear what results one can expect this way, we consider it to be a purer way of calculating recall, with the effect relations coming from a de facto external resource. Table 5.15 shows the results.

The results show a clear trend: EREx has lower recall than OpenIE, while EREx* has a significantly higher recall than OpenIE only when Debatepedia is included in the graph. Importantly, we note that EREx* is only better than EREx in the *full graph* setting. This fits our observation that the effect relations extracted by EREx* tend to be overly specific oftentimes, which is one reason why we proposed the linking to Wikipedia as an additional requirement.

As the recall is particularly low for the settings without Debatepedia, we take a brief look at the few successes in table 5.16:

It is noticeable though unsurprising that the graphs generated with EREx and EREx* contain the exact same test instances. Further, two of them (7,8) are not identified by OpenIE which in turn contains seven instances which EREx and EREx* do not (9–15). One of the latter instances cannot be included in EREx or EREx* because it contains a non-nounphrase as subject (14), but because of the instance’s unspecificity we consider this restriction to be justifiable. The other six of the latter instances could in theory be contained in both EREx and EREx*, but the extractors missed them.

EREx + EREx* + OpenIE	
1	icc $\bar{\rightarrow}$ crimes
2	abortion $\bar{\rightarrow}$ women
3	eating meat $\bar{\rightarrow}$ animals
4	marijuana $\bar{\rightarrow}$ productivity
5	war $\bar{\rightarrow}$ civilians
6	affirmative action $\bar{\rightarrow}$ meritocracy
EREx + EREx*	
7	two-state solution $\bar{\rightarrow}$ stability
8	gay marriage $\bar{\rightarrow}$ procreation

OpenIE only	
9	elections $\bar{\rightarrow}$ judges
10	government $\bar{\rightarrow}$ public transport
11	stimulus $\bar{\rightarrow}$ debt
12	circumcision $\bar{\rightarrow}$ infections
13	primaries $\bar{\rightarrow}$ candidates
14	they $\bar{\rightarrow}$ headaches
15	rights $\bar{\rightarrow}$ contracts

Table 5.16: Effect graph recall (w/o *Debatepedia*): Success analysis.

5.3.3 Explanation Evaluation

As to the best of our knowledge, this is the first approach that generates structured explanations for arguments on an instance-level, we do not compare the results to other systems. The evaluation shows that the method is not potent enough yet and should be seen as a rather conceptual or preliminary approach. We will split the evaluation in two parts: First, we quantify how often we do find explanation candidates and second, we manually analyze the validity of the explanation candidates in relation to their importance ranking.

Data

For our evaluation, we use our dataset containing only arguments from consequences (see chapter 4.3). In order for the proposed explanation types to be applicable, we need to have identified the subject, the verb's polarity, and the object, but not necessarily the sentiment scores and the relation between the subject and the topic. Thus, for simplicity, we do not use StArCon to analyze the argument, but use EREx to extract the effect relation which we then try to explain as discussed previously. Out of the 822 arguments from consequences, we extract 325 effect relations with EREx* and 62 effect relations with EREx. The big difference between the two extractors is of course due to the restrictive linking to Wikipedia in EREx.

We manually annotated for each of these 62 effect relations whether we consider it to make sense. This holds true for 46 of them, while we annotated 6 of them to make at least somewhat sense. For the following evaluation, we focus on these 52 effect relations that make at least somewhat sense, dismissing the 10 non-sense ones. In table 5.17 we

Invalid:			
Holocaust denial opens doors to harmful, non-factual views	Holocaust denial	opens	doors
Somewhat valid:			
Pageants teach kids to follow rules and play fair.	Pageants	teach	kids
Valid:			
Hybrid vehicles reduce noise pollution	Hybrid vehicles	reduce	noise pollution

Table 5.17: Effect relation validity annotation: examples.

present an example for each category. The full annotations are in appendix C.

Quantitative Evaluation

We generate the explanations as described, using the lemmatized effect graph generated by EREx. Overall, we find explanations for only 14 of the effect relations, which leaves 38 effect relations without explanation.

For these 14 effect relations, we find 84 explanation candidates. 57 of them are effect-lexical-explanations and 27 are effect-effect-explanations. For four effect relations, all explanation candidates are effect-lexical-explanations, for three candidates they are all effect-effect-explanations and for the remaining seven there are both effect-lexical- and effect-effect-explanations. Figure 5.3 shows the exact distribution.

Undoubtedly, this shows that with our current approach, we are able to explain at most a small fraction of all the arguments from consequences. However, as we mentioned, we cannot expect that there exists an effect-effect-explanation or effect-lexical-explanation for every argument from consequences. Thus, it remains unclear for now whether the low number of explanation candidates is a shortcoming of the extraction approach or of the explanation type design, or both.

Qualitative Evaluation

For evaluating whether the explanation candidates are valid, we again perform manual annotations into three categories: *invalid*, *somewhat valid*, *valid*. The label *somewhat valid* is used for both explanations that are rather far fetched and also for explanations that are valid for the effect relation but not for the argument they were extracted from.

Overall, we annotated 84 explanation candidates to be bad, 13 ones to be mediocre and only 12 ones to be good. Furthermore, the valid explanations are split among 6 arguments, while for 9 arguments there was at least a mediocre explanation. The anno-

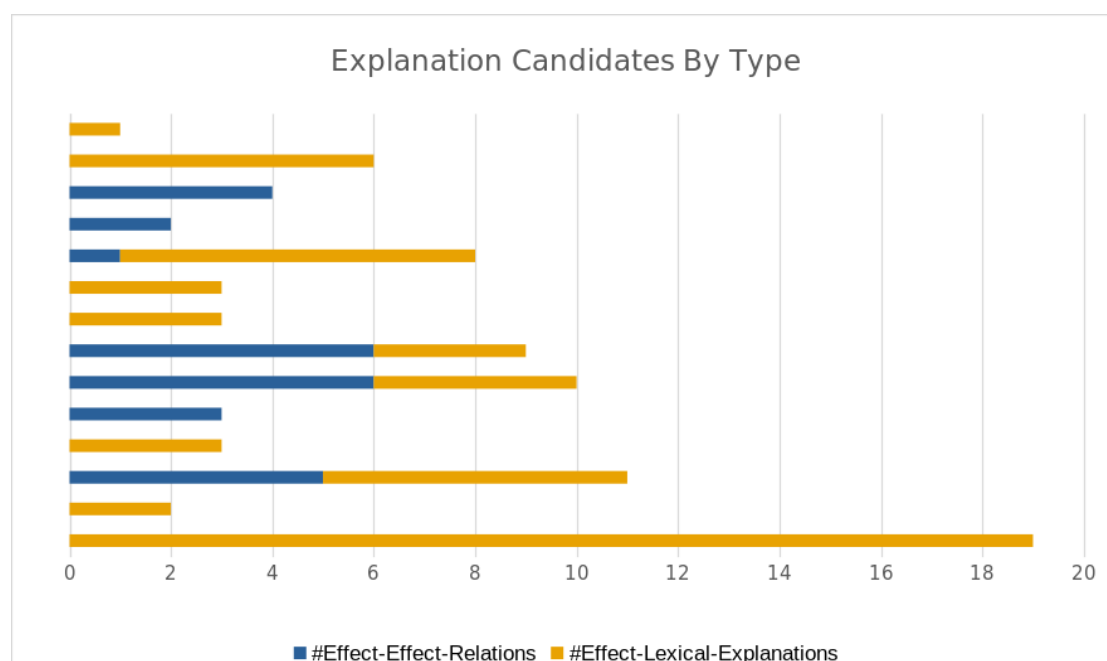


Figure 5.3: Explanation candidates by type: Each bar represents one effect relation. The x-axis shows how many explanation candidates there are for the respective effect relation per explanation type.

tations are presented in appendix D. Figure 5.4 shows the validity distribution among the respective arguments.

Especially for the arguments where explanations candidates of different quality exist, we further check whether our importance measure (see equation 5.3) can help at identifying the valid explanations. We use two different evaluation setups: First, we compute the Spearman Rank Correlation between the importance values and the validity annotation among all the explanation candidates: 0.34. The higher this value, the more expressive is the absolute importance score, which would be especially useful for defining a threshold below which explanation candidates are considered to be invalid. The actual value of 0.34 indicates that the importance score is at least somewhat expressive, but we do not think that it is sufficiently high to identify an exact threshold. However, it might be good enough to define at least a threshold which filters out obvious non-sense. In our limited evaluation data, only one of the 16 explanation candidates with an importance score in the magnitude of 10^{-6} or lower was considered to be somewhat valid while the others were all invalid.

Second, evaluate how well the importance score can be used to rank the explanation candidates for each argument individually. Potentially, the importance score does not generalize well between different arguments, but is better when used for ranking

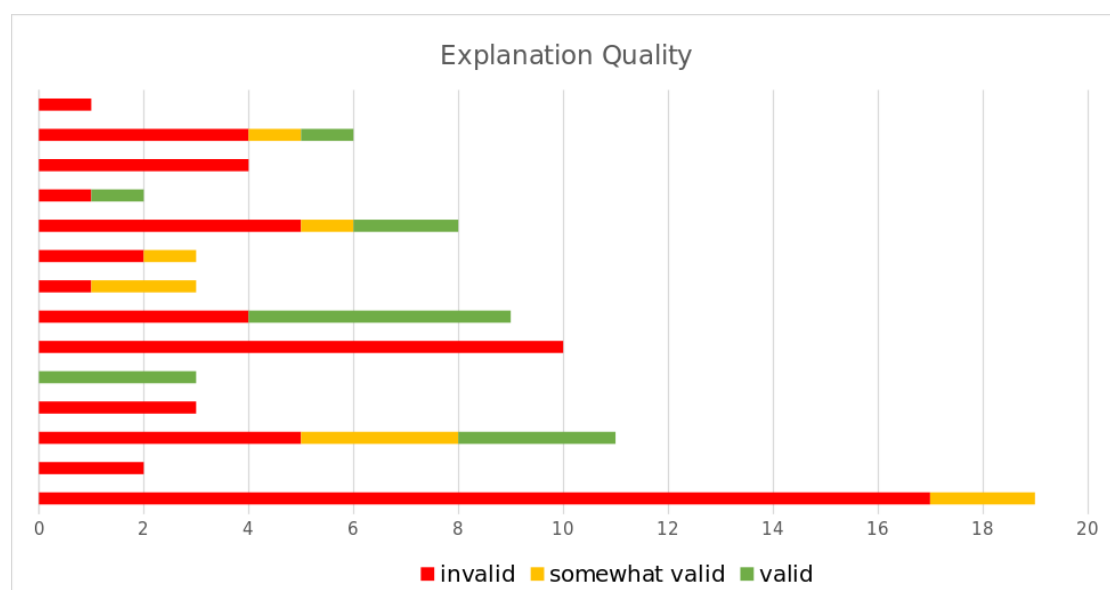


Figure 5.4: Explanation quality per argument. Each bar represents one effect relation. The x-axis shows the amount of explanation candidates and their quality.

<i>Across all arguments</i>	0.343
<i>For each argument individually</i>	
Mean	0.434
Median	0.500
Std deviation	0.496

Table 5.18: Spearman rank correlations between the importance score and the explanation validity.

the explanation candidates obtained for one specific argument. Thus, we compute the Spearman Rank Correlation for every argument specifically. This time, however, we focus on the arguments where explanations of different quality do exist as otherwise, we cannot apply the Spearman Rank Correlation.⁷ The results are shown in Table 5.18. The mean of 0.43 indicates that similarly to our first setting, the importance measure is useful, but needs further improvement. Also, we note that the standard deviation is quite high – this might be partly because our sample size of 7 arguments is very low, but it also shows that the measure completely fails for some arguments: While the rank correlation is positive for six of the arguments, it is -0.48 for one of them. The importance scores and the resulting ranking are included in appendix D.

⁷We also dismiss one argument where three explanation candidates exist which all have the same interim node C and thus the same importance, as in this situation also, computing a Rank Correlation is not possible.

We conclude that we are able to identify only few explanation candidates for few or the arguments, and we are unable to confidently identify the valid explanations. However, the importance measure might still be useful to filter out explanation candidates that are obvious non-sense and it will rank valid explanations higher than invalid ones significantly more often than not.

5.4 Discussion

In this chapter, we proposed a method to extract effect relations from text and used it to build a knowledge graph which we call effect graph. We further proposed a method to use the effect graph as background knowledge for automatically generating structured explanations for arguments from consequences. However, the effect graph’s precision remains unclear while its recall is low. The latter issue might be addressed by either improving the extraction method or, to a certain degree, by running the method on larger text resources. The argument generation works only for few of the arguments used for evaluation. The effect graph can be seen as a valuable resource on its own, as it can potentially be used to also address other tasks than explanation generation, such as identifying (counter-) arguments for a specific topic, or extending common sense KGs like ConceptNet.

While the proposed methods are attractive due to their efficiency, explainability and not needing training data, the limitations are also manifold: The pipeline nature propagates all errors that occur. For instance, the dependency parser in use performs rather poorly on informal texts such as tweets. Further, our definition of positive and negative effect relations is quite shallow and does not always live up to the real world’s complexity. We only capture effect relations that are formulated explicitly within one sentence, and only one effect relation per sentence. Requiring the nodes to link to Wikipedia might be too restrictive while not even truly solving the problem of filtering non-sense nodes. Both the low IAA in our effect graph evaluation as well as the discrepancy of the crowds’ and the expert’s annotations make it hard to assess the correctness of the extracted effect relations. We restrict ourselves to very specific explanation schemes.

It seems natural to extend these explanation schemes by allowing longer paths for effect-effect-explanations or combining effect-effect- and effect-lexical-explanations. However, we do not expect this to work well because one major challenge is to distinguish meaningful explanations from noise and allowing more complex explanations further magnifies this challenge.

What one might consider another limitation is that we do not check the effect relations for factual correctness, which ultimately leads to contradictions and inconsistencies in the effect graph. While fact checking is a difficult and controversial task, we also purposefully decided against any form of fact or consistency checking. Each edge in the effect graph is meant to represent one effect relation exactly as it was expressed. Including critical effect relations in the graph allows for identifying, analyzing, and potentially disproving them.

Chapter 6

Explaining the Judgment Premise: Moral Foundations

¹ In order to explain the judgment premise of an argument from consequences, i.e., to explain why CONSEQUENCE is considered good or bad, we propose to apply the MFT (see chapter 2.1.4). However, instead of specifically addressing CONSEQUENCE's classification, we address the classification of MFs in argumentative texts in general and explore the use of moral framing in online debates.

A debater's moral beliefs go beyond stance and can be expressed with varying sentiment. They play an important role in ideological debates and cannot be resolved by simply comparing facts but often involve a battle of ideas and a clash of different belief systems. Consider the following arguments on whether or not gay marriage should be legal.²

Example 14. *The institution of marriage has traditionally been defined as being between a man and a woman.*

Example 15. *Denying some people the option to marry is discriminatory and creates a second class of citizens.*

Both arguments are based on moral belief systems. The first argument refers to moral values that promote respect for tradition, while the second focuses on fairness

¹This chapter is adapted from Kobbe, J., Rehbein, I., Hulpuş, I., and Stuckenschmidt, H. (2020b). Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics. The corresponding resources are publicly available at <https://github.com/dwslab/Morality-in-Arguments>.

²From <https://gaymarriage.procon.org> (accessed August 25, 2020).

and equal rights. Arguments that express an opposite stance towards the topic usually differ concerning their moral framing. On the other hand, we observe that arguments expressing a similar stance towards a certain topic may still differ with regard to how the argument is framed, as illustrated in the examples 16 and 17 below. While example 16 opposes the legalization of prostitution because it is considered as a harmful form of oppression targeting women, example 17 depicts prostitution as increasing the danger of diseases and contamination. This makes moral framing an interesting ingredient for argument mining.

Example 16. *Prostitution and human trafficking are forms of gender-based violence.*

Example 17. *Prostitution is the biggest vector of sexually transmitted diseases.*

In this chapter, we argue that identifying moral values in debates has the potential to support argument analysis and to help with different subtasks related to argument mining. Being able to distinguish between arguments with similar stance and sentiment but framed according to different moral categories can help to identify new arguments and can improve camp detection, thus supporting more fine-grained modeling of debaters beyond stance. Furthermore, moral framing is of particular interest for the analysis of political debates (Lakoff, 1997; Roggeband and Vliegenthart, 2007).

In practice, however, predicting moral sentiment from text poses several challenges. First, morality is a fuzzy concept, and it is difficult to find an operationalization that turns it into measurable data. Moral sentiment is often expressed implicitly and thus hard to detect, based merely on the presence of lexical cues. In addition, human coders might be biased by their own belief systems, which casts doubt on the validity of the annotations used to train or evaluate automatic systems.

In this chapter, we present an evaluation of different models for the prediction of moral framing in text on two datasets, one of which consists of arguments, and assess the benefits of these predictions for the analysis of arguments. Based on three datasets with argumentative text, we investigate whether we can find correlations between moral values and different aspects of argumentation, such as argument quality, stance, or audience approval.

Our main contributions are the following: (i) We augment the ArgQuality Corpus of Wachsmuth et al. (2017a) with annotations for moral values, as a first test set for the evaluation of moral sentiment in argumentation; (ii) We evaluate two methods for the prediction of moral sentiment on the new dataset; (iii) We present a correlation study investigating the relation between moral framing and argument quality, stance, and audience reactions.

	<i>Care</i>	<i>Fairness</i>	<i>Loyalty</i>	<i>Authority</i>	<i>Purity</i>	<i>Moral</i>
Cohen's κ	.469	.407	.529	.363	.280	.434
Krippendorff's α	.459	.400	.530	.356	.255	.402
Absolute Agreement (pos/neg)	60/187	16/267	10/294	12/274	13/257	165/68
Absolute Disagreement	73	37	16	34	50	87

Table 6.1: Inter-Annotator Agreement for the five MFs and for a binary label (*Moral*: yes/no).

The chapter is structured as follows. In section 6.1, we describe the annotation of our test set and present different approaches to the automatic detection of moral sentiment in debates. In section 6.2, we present our correlation analysis. We conclude the chapter with a discussion in section 6.3.

6.1 Predicting Moral Sentiment in Tweets and Debates

6.1.1 A New Test Set for Moral Framing in Argumentation

As a test set for evaluating moral framing in English argumentative text, we use the Dagstuhl ArgQuality Corpus Wachsmuth et al. (2017a). The dataset contains 320 arguments with approx. 22,600 tokens, covering 16 topics, and is balanced for stance. The data was extracted from two online debate platforms by Habernal and Gurevych (2016). Each instance has been annotated by three coders, using a fine-grained scheme to assess the arguments' quality. The data also provides a majority score for each dimension of argument quality (Wachsmuth et al., 2017a). The authors report a low agreement for the individual annotations (.51 Krippendorff's α) but a high majority agreement (94%).

We further augment the ArgQuality corpus with annotations for moral foundations, manually coded by two of the authors. We chose not to annotate the 10 MTs encoded in the dictionary but considered the two ends of each dimension (virtue/vice) as one category (MF). The motivation behind this decision is that the two MTs of an MF are closely related, and it is often unclear which end of the dimension is addressed, particularly for negated sentences. E.g., "I could never hurt you" could either be considered as an instance of Harm as it uses vocabulary related to this dimension or could be annotated as the opposite, Care, as it talks about *not* being able to harm somebody, thus being more strongly related to the *virtue* class.

Table 6.1 shows IAA scores for individual MFs on the ArgQuality dataset. As expected, IAA is low, being roughly in the same range as agreement scores reported for the annotation of emotions (Schuff et al., 2017; Wood et al., 2018), thus giving evidence for

the subjectivity of the task. Our IAA is not directly comparable to Hoover et al. (2020) as they report Fleiss' κ for the 10 MTs, with an avg. of .315 κ over all 10 classes.

6.1.2 Methods for the Prediction of Moral Sentiment

We model moral sentiment prediction as a text classification task and propose two distinct methods for feature generation. The first method is based on a *sense-disambiguated* version of the MFD and extends its coverage by exploiting relations in WordNet. The second method uses the MFD as seed data to learn BERT sequence embeddings that encode moral sentiment. The representations created by each method are fixed-sized vectors that can easily be combined by concatenation.

Sense-disambiguated features (WN-PPR) The MFD has two main disadvantages that we try to overcome with this method. First, the lexicon contains many words with different word senses, where the moral value only applies to one specific sense. Thus, we link the dictionary entries to their corresponding WordNet synsets. This way, *fair* is only considered to be related to the MF *fairness-cheating* if used as synonym to *just* or *honest*, but not if used as synonym to *carnival*, *funfair* or *attractively feminine*. Also, this way we overcome the problems resulting from the use of regular expressions in the MFD (e.g., *defenestration* would belong to the MF *Care* because of the entry *defen**, and *Churchill* would trigger *Purity* because of the entry *church**). The second disadvantage of the MFD is its low coverage, which we extend by running PPR (Haveliwala, 2003) on the set of WordNet synsets that have been linked to dictionary entries (see also Hulpus et al. (2020)).

Linking MFD entries to WordNet To create a word sense disambiguated version of the MFD, one expert annotator was presented with the following information: (i) a specific MT; (ii) a WordNet synset whereof at least one word in the synset is part of the respective MT in the MFD; and (iii) its definition. With this information, the annotator decided whether the synset is relevant for the moral foundation in question or not. Overall, the resulting lexicon contains, on average, 61 synsets per MF.

Extending the disambiguated lexicon We extend the disambiguated lexicon by exploiting relations between synsets in WordNet, such as *hypernym* or *similar to*. Concretely, we run PPR on the graph consisting of the WordNet synsets and the relations between them for every MF, using the corresponding lexicon entries as seed nodes. This

way, each WordNet synset is assigned a fixed-sized vector containing scores for each MF, including the category *GeneralMorality*. *GeneralMorality* includes terms related to moral concepts that do not fit into one of the five MFs, like *ethic*, *good*, *evil*. We expect that higher scores reflect a stronger correspondence between the synset and the respective MF.

Extracting features from text Given a short English text, we first extract WordNet 3.0 synsets using the disambiguation method by Tan (2014). Then we link these synsets to WordNet 3.1, using the official WordNet Search Engine³ and, if necessary, resolving the final mapping manually. For instance, a variety of offensive terms have been removed in WordNet 3.1, and thus, we had to link terms like *darky* or *tom* to *black (noun.person)* manually.⁴ As each of the synsets is assigned a fixed-size score vector in our lexicon, any function to aggregate these vectors is conceivable. To obtain vectors that do not depend on the input text’s length, we decided to take their mean. The result is a vector consisting of six entries, where each entry represents a MF, including *GeneralMorality*.

Contextualized MF sequence embeddings (SBERT-Wiki) Our second method uses Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to obtain text representations that encode moral sentiment. We fine-tune SBERT embeddings so that they encode different moral foundations. First, we download all short Wikipedia abstracts from DBpedia and label them with their corresponding MF (if any), using weak supervision. Our approach is based on the MFD and proceeds as follows: For each dictionary entry, we search in Wikipedia for corresponding articles to get a set of candidates consisting of articles whose title is a lexicon entry (including redirections) and articles that are linked by the lexical entry’s disambiguation page. From these candidates, we manually select the ones related to the MF and label their abstracts accordingly.

This approach yields 317 short abstracts from Wikipedia, labeled with moral foundations, extracted from a pool of 4,935,596 unlabelled short Wikipedia abstracts. We iterate over each abstract in the annotated dataset, considering the abstract as the anchor text. First, we retrieve all other abstracts labeled with the same moral foundation as the anchor and create pairs of (anchor, positive sample). Then, for each pair, we ran-

³<http://wordnet-rdf.princeton.edu/json/pwn30/...>

⁴We are aware that this treatment is not optimal. A better solution would link those terms to a synset that captures their offensive usage, similar to the one for *Kraut*: *offensive term for a person of German descent*.

domly select 3 labeled abstracts that belong to a different moral foundation as well as 7 abstracts from the unlabelled pool as negative samples, assuming that the unlabelled abstracts also do, more often than not, either belong to a different moral foundation or do not express any moral content. This gives us a total of 10 negative samples for each pair and results in a weakly supervised dataset with 107,940 instances. We then fine-tune the model on the data, using the same settings as reported in Reimers and Gurevych (2019). After the training is completed, we use the learned model to retrieve representations for new text sequences from different argumentation datasets, expecting that the fine-tuned embeddings will now capture some aspects of moral sentiment. We compare our approach with the pretrained SBERT embeddings (bert-base-nli-stsb-mean-tokens) of Reimers and Gurevych (2019), trained without the fine-tuning step on the weakly supervised Wikipedia abstracts.

6.1.3 State of the Art and Baselines

multi-label BERT To compare our lexicon-based methods with a state-of-the-art approach to text classification, we train a multi-label text classifier based on BERT. We use a publicly available implementation in pytorch⁵ that replaces the cross-entropy loss with a binary cross-entropy with logits to adapt the BERT sequence classifier to the multi-label setup.

The model includes an input embedding layer for the pretrained BERT embeddings, the BERT encoder with 12 attention layers, and, as final layer, a linear transformation, with one dimension for each class. This gives us six output dimensions: the five moral foundations + one class for tweets with non-moral content. Our model uses the pretrained English uncased BERT base embeddings with a vocabulary size of around 30,000. We use the same data splits and preprocessing in all experiments (see section 6.1.4). In contrast to our other models, however, BERT further segments the input text into subword tokens (WordPiece tokenization), which might increase coverage for words not seen in the training data.

Random baseline The *Random* baseline assigns labels randomly but respecting the class distribution in the training data. Results are averaged over 100 trials.

⁵The code was adapted from <https://medium.com/huggingface/multi-label-text-classification-using-bert-the-mighty-transformer-69714fa3fb3d> and is based on the HuggingFace library (<https://github.com/huggingface/pytorch-pretrained-BERT>).

MFD baseline Given a text, we compute frequency counts for each MF, based on the entries in the MFD, and normalize by text length. We use these count-based vectors as features for the text classifier. Similar to WN-PPR, we derive one feature per MF, including general morality.

6.1.4 Data

We now present the data used for the evaluation of the methods described in section 6.1.2 for the prediction of moral sentiment in tweets and debates. As training data for our MF classifiers, we use the Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020), a collection of approximately 35,000 tweets covering seven controversial topical threads: *All Lives Matter*, *Black Lives Matter*, *the Baltimore protests*, *the 2016 Presidential election*, *hate speech & offensive language* (Davidson et al., 2017), *Hurricane Sandy*, and *#MeToo*. Each tweet has been annotated with MFs by at least three trained annotators. The authors report Fleiss’ κ and PABAK, a measure adjusted for prevalence and bias (Sim and Wright, 2005). IAA is relatively low (with a Fleiss κ in the range of 0.24 – 0.46 and PABAK ranging from 0.65 – 0.85) and shows considerable variation across the different moral domains and threads.

We follow the procedure described in Hoover et al. (2020) to create a gold standard from the annotated tweets and consider a label as *gold* if it was assigned by at least half of the annotators. Thus, our gold standard includes 6 labels: one for each MF and a sixth one for GeneralMorality. Note that in the MFTC, this label is called *Non-moral* while we report results for its inverse, which we call *Moral*. We normalized the tweets using the script available from the Glove website.⁶ We noticed that the dataset includes many near-duplicates (e.g., 96 instances of *homosexuality is a sin*). To ensure that these near-duplicates do not appear in both training *and* test set, we split the data into the different threads and present results for a seven-fold cross-validation where we train the models on six threads and evaluate on the remaining one. We also evaluate the models trained on the MFTC on out-of-domain data from the ArgQuality Corpus, where we consider all labels assigned by each of the two annotators as ground truth.⁷

⁶<https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb>

⁷As the data has been annotated by two of the authors, we can be sure that we do not have to eliminate spammers.

Method	Moral	Care	Fairness	Loyalty	Authority	Purity	Average
<i>Random baseline</i>	.519	.173	.169	.100	.099	.055	.119
<i>MFD baseline</i>	.630	.332	.213	.166	.231	.141	.217
<i>multi-label BERT</i>	.669	.510	.573	.437	.377	.363	.452
<i>WN-PPR</i>	.628	.334	.379	.311	.210	.088	.264
<i>SBERT-Base</i>	.685	.434	.511	.372	.327	.214	.372
<i>SBERT-Wiki</i>	.697	.463	.516	.377	.341	.220	.383
<i>WN-PPR + SBERT-Wiki</i>	.689	.446	.520	.387	.346	.230	.386

Table 6.2: Binary F1-scores on the MFTC for individual MFs (F1 for the positive class). The last column shows the average over the F1 scores for the five MFs (excluding *Moral*).

Method	Moral	Care	Fairness	Loyalty	Authority	Purity	Average
<i>Random baseline</i>	.658	.257	.179	.096	.134	.105	.154
<i>MFD baseline</i>	.853	.056	.237	.043	.200	.086	.124
<i>multi-label BERT</i>	.444	.517	.519	.138	.157	.208	.308
<i>WN-PPR</i>	.756	.118	.253	.049	.105	.029	.111
<i>SBERT-Base</i>	.703	.280	.342	.065	.148	.133	.194
<i>SBERT-Wiki</i>	.730	.339	.246	.125	.233	.318	.252
<i>WN-PPR + SBERT-Wiki</i>	.686	.298	.351	.067	.040	.135	.178

Table 6.3: Binary F1-scores on the Dagstuhl ArgQuality Corpus for individual MFs (F1 for the positive class). The last column shows the average over the F1 scores for the five MFs (excluding *Moral*).

6.1.5 Results for MF Prediction on Tweets and Debates

We conduct experiments on the Twitter corpus, testing different traditional classification methods, and report results for the best performing classifiers only. For WN-PPR and MFD-Features, this was a k-nearest-neighbors classifier, and for SBERT-Base, SBERT-Wiki, as well as for WN-PPR + SBERT-Wiki a Linear Discriminant Analysis.⁸ All other results, as well as the correlations reported in section 6.2, refer to these classification methods.

Table 6.2 shows results on the MFTC for our different methods. Not surprisingly, multi-label BERT outperforms all other methods on the Twitter data. However, our lexicon-based methods outperform the random baseline for each category, with the best results obtained by the concatenation of SBERT-Wiki with WN-PPR. WN-PPR on its own only yields poor results, barely outperforming the constant and the MFD baseline.

When applying the classifiers to the out-of-domain data from the ArgQuality corpus (Table 6.3), multi-label BERT still yields best results, but now SBERT-Wiki outperforms BERT on the *Authority*, *Purity* and *Moral* categories. The lower performance for

⁸We use the scikit-learn implementation for these methods. Other methods we tried include logistic regression, decision trees, naïve bayes, support vector machines.

the *Moral* class can be explained by the differences in class distribution between the two datasets. In the MFTC, this class makes up for approximately 57% of the training instances, while the amount of moral instances in the debate corpus is much higher (79%). The lexicon-based methods are not sensitive to the class distribution in the training data, which, in this case, makes them more robust. Still, all systems fail to beat the majority baseline for the Moral class which has a binary F1-score of 0.881.⁹ WN-PPR again performs poorly with results below the random baseline, and results for the MFD baseline also fail to outperform the random baseline. This time, results for the concatenation of WN-PPR and BERT-Wiki are considerably worse than for BERT-Wiki alone.

6.2 Correlation Studies

To study the impact of moral framing in argumentation, we investigate the correlation between moral sentiment and other properties of argumentative text, namely argument quality, stance, and audience reactions. For this, we use the multi-label BERT model that yielded the best results on both datasets.

6.2.1 Data

The **Dagstuhl ArgQuality Corpus** contains arguments that are annotated with different dimensions of argument quality, such as *cogency* and *credibility*, as well as a score for *overall quality*. Some of the dimensions are also interesting for contexts other than argument quality, such as *clarity* and *emotional appeal*.

The **IBM Argument Quality Ranking Corpus** (Gretz et al., 2019) is used to triangulate our findings on the Dagstuhl ArgQuality Corpus and to investigate the correlation between moral sentiment and an argument’s stance. The corpus contains more than 30,000 arguments on 71 topics, labelled for quality (*good* or *bad*) and stance (*pro* or *con*) by crowd annotators. To obtain ranks for argument quality, the authors apply two different strategies, which both give more weight to the answers of reliable annotators.

We use **CORPS** (Guerini et al., 2013) to investigate whether moral sentiment in political speeches has an impact on the audience. CORPS includes >3,600 political speeches held by more than 203 different speakers, tagged for audience reactions such as *applause*, *laughter* or *booing*. The motivation for creating the corpus was that such tags might highlight passages in the speech where an attempt has been made by the

⁹The majority baseline is not included in tables 6.2 and 6.3 because its binary F1-score is zero for all classes except *Moral*.

	Care		Fairness		Loyalty		Authority		Purity		Moral	
	HU	BM	HU	BM	HU	BM	HU	BM	HU	BM	HU	BM
<i>Dagstuhl ArgQuality Corpus</i>												
overall quality	.25	.15	.10	.08	.05	.10⁻	-.09	.05 ⁺	.03	.07 ⁻	.19	.21
local acceptability	.18	.09	.00	-.04 ⁺	.00	.04 ⁻	-.15	-.01 ⁺	-.03	.07 ⁻	.03	.09
appropriateness	.30	.17	-.01	.03	-.02	.05 ⁻	-.09	.00 ⁺	.01	.02	.19	.15
arrangement	.24	.16	.08	.03	.03	.08 ⁻	-.06	-.01 ⁺	.04	.05	.16	.17
clarity	.17	.17	.02	.02	.05	.12⁻	-.03	-.01 ⁺	.03	.06	.09	.21⁻
cogency	.24	.16	.05	.06	-.02	.03 ⁻	-.10	.05 ⁺	.01	.03 ⁻	.10	.18
effectiveness	.25	.17	.09	.09	-.05	.07 ⁻	-.10	-.02 ⁺	.04	.04	.21	.17
global acceptability	.23	.12	.05	.05	-.01	.07 ⁻	-.12	.02 ⁺	.01	.04	.12	.13
global relevance	.15	.06	.11	.09	.02	.07 ⁻	-.11	.00 ⁺	.04	.05	.12	.11
global sufficiency	.19	.11	.11	.11	-.01	.06 ⁻	-.04	-.03 ⁺	.07	.05 ⁻	.19	.14⁺
reasonableness	.23	.17	.09	.08	.02	.08 ⁻	-.11	.04	.06	.07 ⁻	.16	.18
local relevance	.18	.14	.08	.03	.01	.01 ⁻	-.10	.02 ⁺	.02	-.02	.12	.13
credibility	.22	.07	.06	.02 ⁻	.05	-.01	-.13	.01 ⁺	-.01	.00 ⁻	.09	.08
emotional appeal	.32	.22	.16	.12⁺	.14	.02 ⁻	-.01	.10 ⁺	-.01	.02	.31	.25
sufficiency	.25	.18	.06	.09 ⁻	.00	.04 ⁻	-.10	.03 ⁺	.07	.06 ⁻	.15	.19⁺
<i>IBM-AQR</i>												
quality (WA)	.08		.06		.01		.00		-.02		.08⁻	
quality (MACE-P)	.08		.05		.01		.00		-.01		.07⁻	
stance	.07		.01		-.03		.01		-.03⁺		.04	
<i>CORPS</i>												
applause	.02		.04		.07		.05		.01		.10	
laughter	-.07		-.05		-.05		-.03		-.02		-.11	

Table 6.4: Spearman ρ between human annotations (HU) and multi-label BERT predictions (BM), respectively, and quality, stance and audience reactions. Bold values are statistically significant ($p < 0.05$). ⁺/₋ : The correlation to the SBERT-Wiki predictions was considerably higher / lower (by at least 0.05).

speaker to persuade the audience, either successful or not. We expect to find a correlation between text passages that triggered a positive audience reaction (i.e. *applause*) and moral framing, but not for *laughter*. We focus on these two tags as the other tags are relatively rare in comparison.¹⁰ We also exclude mixed tags that mark two different reactions for the same text passage (*laughter; applause*). To test our hypothesis, we predict moral sentiment for the speech passages directly before an audience reaction was triggered. We consider up to 360 tokens of speech context and omit all speech passages where another tag occurs within this context.

6.2.2 Results for the Correlation Analysis

Table 6.4 shows Spearman correlations between argument quality, stance, and audience reactions and a) human annotations (HU) and b) labels predicted by multi-label BERT

¹⁰ Applause: 23,095; Laughter: 5,857; Booming: 532; Cheers: 80; Sustained applause: 61; Spontaneous demonstration: 16.

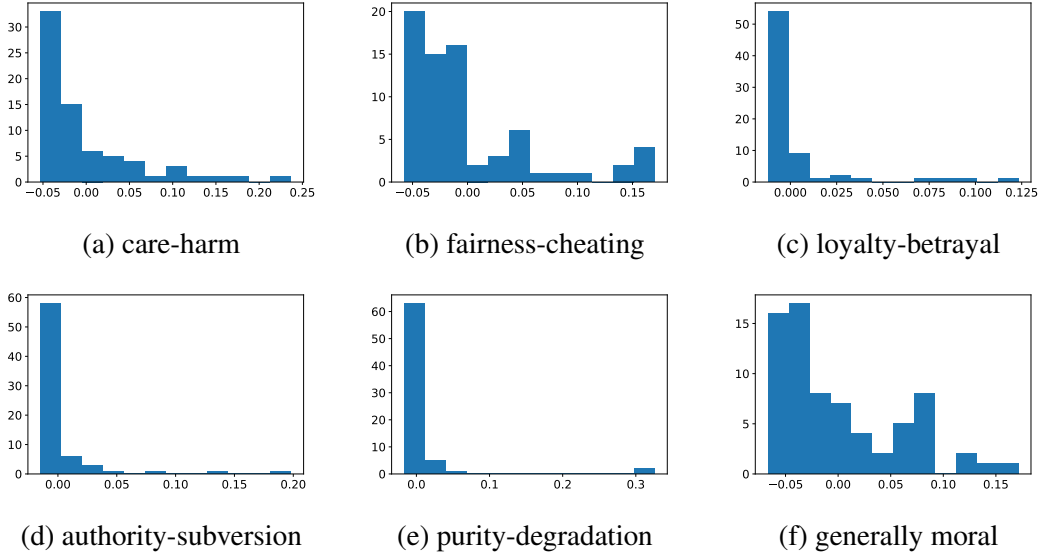


Figure 6.1: Histograms of Spearman correlations between topics and MFs (multi-label BERT).

(BM). We observe a weak positive correlation between argument quality and moral sentiment for the two most frequent categories (*Moral*, *Care*) on the ArgQuality data. For the other MFs, there are no significant effects. On the IBM-AQR Corpus, we see a consistent and significant positive correlation for *Care* and *Fairness*. However, the effect is very weak. For the subdimensions of argument quality, the correlations tend to be similar to the ones for overall quality and are highest for *emotional appeal*, which seems plausible. Concerning argument stance, we again find slightly positive correlations for *Care* and *Moral*. Results on the CORPS data are as expected: a positive correlation for *applause* and a negative one for *laughter*, but again the effect is very weak.

Correlation with topic To control for topic effects, we computed the correlation between topic and argument quality, between topic and stance, and between topic and MF in the IBM-AQR. While we found no correlation between topic and argument quality or stance, there was a weak correlation between some topics and specific MFs.

Figure 6.1 shows the distribution of correlations between topics and the predicted MFs. For most topics, the correlation for certain MFs is slightly negative or close to 0, but there are very few topics that have a relatively high correlation to certain MFs. The concrete topics having correlations whose absolute value is higher than 0.16 are shown in table 6.5.

<i>care-harm</i>	<i>fairness-cheating</i>	<i>purity-degradation</i>
We should ban targeted killing Assisted suicide should be a criminal offence We should fight for the abolition of nuclear weapons	The use of public defenders should be mandatory We should end affirmative action We should abolish intellectual property rights	The vow of celibacy should be abandoned We should prohibit school prayer
<i>loyalty-betrayal</i>	<i>authority-subversion</i>	<i>generally moral</i>
	We should prohibit flag burning	We should ban targeted killing

Table 6.5: Topics having a Spearman correlation higher than 0.16 to MF (muti-label BERT), ordered descending per MF.

6.3 Discussion

We evaluated different models for predicting moral sentiment in debates, based on the MFT. We then used our models to predict moral values in three argumentation datasets. We investigated whether we could find a correlation between morality and (i) argument quality, (ii) stance, and (iii) audience reactions for political speeches.

We found weak but significant correlations between general morality and argument quality in the ArgQuality data and a consistent positive correlation between moral sentiment and audience approval in CORPS as well as a negative correlation for moral sentiment and laughter. However, our study has several limitations that need to be addressed. One problem is the low accuracy of the classifiers for the prediction of moral values. While results were substantially higher than the random baseline and an MFD-based baseline, we still expect a considerable amount of noise in the classifiers' predictions, which might impact the results of the correlation study. It is conceivable that cleaner predictions might increase the effect size of the observed correlations, which would be consistent with the slightly larger correlation coefficients found for the human annotations. This, however, still needs to be confirmed.

A crucial issue for reliably classifying MFs concerns the reliability of the human annotations. While we expect that more extensive training and more detailed guidelines will increase IAA for human annotation at least slightly, we still think that due to the fuzziness of the concept of morality, high agreement scores are not very probable. Thus, we would like to propose a different approach to the annotation of MFs where we ground the annotations in lexical semantics. This approach has already been shown to improve IAA for a similarly difficult annotation task, namely the annotation of causal language (Dunietz et al., 2015). The authors created a lexical resource for terms that can trigger

causality in text and instructed annotators to disambiguate instances of those terms in context, showing that their modularized, dictionary-based approach yields substantially increased IAA scores.

Being able to predict moral values in text reliably can open up new research avenues in argumentation. E.g., recent work in psychology has shown that moral values play an important role in debates on political and social issues (Feinberg and Willer, 2013; Voelkel and Feinberg, 2018; Feinberg and Willer, 2019). For example, Feinberg and Willer (2013) have shown that debates on environmental issues are often framed in terms of moral values such as *Care-Harm*, a MF that is at the core of liberal belief systems, while conservatives, in contrast, seem to value all five MFs more similarly (Graham et al., 2009). This often results in highly polarized discussions, and Feinberg and Willer (2013) argue that reframing such issues in terms of moral values that explicitly address the opponents’ belief system might have the potential to depolarize controversial debates and improve understanding between the camps by addressing the “moral empathy gap” (see Feinberg and Willer (2019) and references therein).

We consider the classification of MFs in texts, especially arguments, to be a step towards our goal of explaining why the consequence of an argument from consequences is considered good or bad. However, we also experimented with an alternative approach which is more in line with our work in chapter 5. In Hulpuş et al. (2020), we propose a method to project MFs on KGs. Starting with the MFD, we manually link its entries to existing KGs (DBpedia, ConceptNet, WordNet) and thereby disambiguate the MFD entries. Then, using a variant of PPR, we score the remaining entities in the KG with respect to the moral dimensions (MFs, MTs and *virtue vs vice*). This work nicely complements our proposed method for explaining the effect premise where CONSEQUENCE is represented in a knowledge graph, the effect graph.

Chapter 7

Conclusion

Arguments from consequences are frequently used in online debates. Although being defeasible, they follow a clear scheme which we reconstruct and use as a basis to explain a given argument. Our scheme reconstruction involves concretely identifying ACTION and its CONSEQUENCE. Our proposed method evolves around effect verbs and is intuitive and unsupervised. By identifying CONSEQUENCE’s alignment using simple rules involving an effect and a sentiment lexicon, we can infer the argument’s conclusion. We evaluate the proposed method by addressing the stance detection task. While outperforming modern transformer based models in some settings, our method generally is inferior to them in terms of stance detection, but offers the advantage of needing no training and being explainable by providing the effect triples.

For explaining the effect premise of an argument from consequences, we build a knowledge graph which we call effect graph. It contains effect relations which we use as background knowledge to explain the effect relation at the core of the effect premise. The underlying intuition is that effect relations are often transitive and thus an effect can be explained by two effect relations. Additionally, we include lexical knowledge in the proposed explanations. While the task of explaining arguments is difficult to evaluate since most often there exist different explanatory approaches, we consider our results to be promising although still far from truly solving the task. Further, with the effect graph, we provide a resource which might be interesting also in other use cases such as querying for specific arguments or extending common sense KGs like ConceptNet.

Lastly, for explaining the judgment premise of an argument from consequences, we examine the use of the MFT for analyzing arguments. While there exists a dictionary for terms related to certain MFs, it is both sparse and ambiguous. Thus, we extended and disambiguated the dictionary using WordNet and trained MF classifiers either based

on the extended dictionary or pretrained language models. We performed a correlation study with the best performing classifier and found positive correlations between the reference to MFs and argument quality as well as audience reactions. However, a reliable classification of MFs is not possible yet. In our opinion, the main problem is in the operationalization of the MFs, as indicated by the generally weak IAA.

Altogether, despite many limitations, we addressed the analysis of arguments from consequences in detail and, being provided an argument from consequences consisting of the effect premise, propose methods to (i) identify the scheme components ACTION and CONSEQUENCE; (ii) classify whether CONSEQUENCE is considered good or bad; (iii) reconstruct the judgment premise and the conclusion; (iv) predict the argument's stance towards a given topic; (v) offer structured explanations for why ACTION leads to CONSEQUENCE; (vi) relate the argument and the judgment premise in particular to MFs.

On a higher level, we consider our work to be a step towards knowledge enhanced modeling of arguments and debates. While modern generative pretrained transformers can provide explanations in natural language with astounding quality, they are unable of real reasoning yet. Modeling arguments in a modular way and considering the arguments' (informal) logic is an important complement to such language models which enables deeper analyses of arguments and debates, as well as improved reasoning capabilities.

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project ExpLAIN, Grant Number STU 266/14-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

Bibliography

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.
- Addawood, A., Schneider, J., and Bashir, M. (2017). Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th International Conference on Social Media & Society, #SMSociety17*, pages 2:1–2:10, New York, NY, USA. ACM.
- Afantenos, S., Peldszus, A., and Stede, M. (2018). Comparing decoding mechanisms for parsing argumentative structures. *Argument & Computation*, 9(3):177–192.
- Al-Khatib, K., Hou, Y., Wachsmuth, H., Jochim, C., Bonin, F., and Stein, B. (2020). End-to-end argumentation knowledge graph construction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7367–7374.
- Al Khatib, K., Trautner, L., Wachsmuth, H., Hou, Y., and Stein, B. (2021). Employing argumentation knowledge graphs for neural argument generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4744–4754, Online. Association for Computational Linguistics.
- Al-Thanyyan, S. S. and Azmi, A. M. (2021). Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.
- Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9. Association for Computational Linguistics.
- Angeli, G., Premkumar, M. J. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

- Araque, O., Gatti, L., and Kalimeri, K. (2020). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Bach, S. H., Huang, B., London, B., and Getoor, L. (2013). Hinge-loss Markov random fields: Convex inference for structured prediction. In *Conference on Uncertainty in Artificial Intelligence*.
- Badaro, G., Jundi, H., Hajj, H., and El-Hajj, W. (2018). Emowordnet: Automatic expansion of emotion lexicon using english wordnet. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 86–93.
- Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017a). Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.
- Bar-Haim, R., Edelstein, L., Jochim, C., and Slonim, N. (2017b). Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics.
- Becker, M., Hulpuş, I., Opitz, J., Paul, D., Kobbe, J., Stuckenschmidt, H., and Frank, A. (2020). Explaining arguments with background knowledge: Towards knowledge-based argumentation analysis. *Datenbank-Spektrum*, 20:131–141.
- Bevilacqua, M. and Navigli, R. (2020). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Black, E. and Hunter, A. (2012). A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, 22(1):55–78.
- Boltužić, F. and Šnajder, J. (2016). Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133. Association for Computational Linguistics.
- Bunescu, R. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation.

- Burke, M. B. (1985). Unstated premises. *Informal logic*, 7(2).
- Burnyeat, M. F. (1994). Enthymeme: Aristotle on the logic of persuasion. In *Aristotle's Rhetoric*, pages 3–56. Princeton University Press.
- Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21.
- Cheng, A.-S. and Fleischmann, K. R. (2010). Developing a Meta-inventory of Human Values. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 3:1–3:10, Silver Springs, MD, USA. American Society for Information Science. event-place: Pittsburgh, Pennsylvania.
- Chinsha, T. and Joseph, S. (2015). A syntactic approach for aspect based opinion mining. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)*, pages 24–31. IEEE.
- Choi, Y. and Wiebe, J. (2014). +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191. Association for Computational Linguistics.
- Corro, L. D. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.
- Davidson, D. (1967). Causal relations. *The Journal of Philosophy*, 64(21):691.
- Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.
- De Marneffe, M.-C. and Manning, C. D. (2008). The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dey, K., Shrivastava, R., and Kaushik, S. (2018). Topical stance detection for twitter: A two-phase lstm model using attention. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Advances in Information Retrieval*, pages 529–536. Springer International Publishing.
- Du, J., Xu, R., He, Y., and Gui, L. (2017). Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.
- Dunietz, J., Levin, L., and Carbonell, J. (2015). Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196, Denver, Colorado, USA. Association for Computational Linguistics.
- Durmus, E. and Cardie, C. (2018). Exploring the role of prior beliefs for argument persuasion. 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 1035–1045. Association for Computational Linguistics.
- Durmus, E. and Cardie, C. (2019). A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Dusmanu, M., Cabrio, E., and Villata, S. (2017). Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.
- Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.
- Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. pages 174–179.
- Federici, M. and Dragoni, M. (2016). A Knowledge-Based Approach for Aspect-Based Opinion Mining. In *SemWebEval@ESWC*.
- Feinberg, M. and Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science*, 24(1):56–62. PMID: 23228937.
- Feinberg, M. and Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social Psychology and Personality Compass*, pages 56–62.
- Fellbaum, C. (2010). Princeton university: About wordnet.
- Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In *ACL*.
- Fischer, D. H. (1971). *Historians’ Fallacies : Toward a Logic of Historical Thought*. Routledge & Kegan Paul.

- Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., and Dehgani, M. (2019). Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*.
- Fulgoni, D., Carpenter, J., Ungar, L., and Preoțiuc-Pietro, D. (2016). An Empirical Exploration of Moral Foundations Theory in Partisan News Sources. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *The 10th International Conference on Language Resources and Evaluation, LREC'16*, pages 3730–3736, Paris, France. European Language Resources Association (ELRA).
- Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., and Dehghani, M. (2016). Morality Between the Lines: Detecting Moral Sentiment In Text. In *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*.
- Gashteovski, K., Gemulla, R., and Del Corro, L. (2017). MinIE: Minimizing Facts in Open Information Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640. Association for Computational Linguistics. event-place: Copenhagen, Denmark.
- Ghosh, S., Anand, K., Rajanala, S., Reddy, A. B., and Singh, M. (2018). Unsupervised Stance Classification in Online Debates. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, CoDS-COMAD '18*, pages 30–36, New York, NY, USA. ACM. event-place: Goa, India.
- Ghosh, S., Singhanian, P., Singh, S., Rudra, K., and Ghosh, S. (2019). Stance detection in web and social media: A comparative study. In Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D. E., Heinatz Bürki, G., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87. Springer International Publishing.
- Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12, MultiSumQA '03*, page 76–83, USA. Association for Computational Linguistics.
- Girju, R. and Moldovan, D. (2002). Text mining for causal relations. In *FLAIRS conference*, pages 360–364.
- Gough, J. and Tindale, C. W. (1985). "hidden" or "missing" premises. *Informal Logic*, 7(2).
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, 47:55 – 130.
- Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5:1029).

- Green, N. L. (2018). Towards mining scientific discourse using argumentation schemes. *Argument & Computation*, 9(2):121–135.
- Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2019). A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*.
- Guerini, M., Giampiccolo, D., Moretti, G., Sprugnoli, R., and Strapparava, C. (2013). The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions. In Poggi, I., D’Errico, F., Vincze, L., and Vinciarelli, A., editors, *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Habernal, I. and Gurevych, I. (2016). What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.
- Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.
- Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.
- Haidt, J. and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.
- Hasan, K. S. and Ng, V. (2013). Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–821. Association for Computational Linguistics.
- Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web, WWW ’02*, page 517–526, New York, NY, USA. Association for Computing Machinery.
- Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.
- He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what’s next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

- He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention.
- Heindorf, S., Scholten, Y., Wachsmuth, H., Ngonga Ngomo, A.-C., and Potthast, M. (2020). Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, page 3023–3030, New York, NY, USA. Association for Computing Machinery.
- Hitchcock, D. and Hitchcock, D. (2017). Enthymematic arguments. *On Reasoning and Argument: Essays in Informal Logic and on Critical Thinking*, pages 39–56.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hogan, A., Blomqvist, E., Cochez, M., D’amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, 54(4).
- Hoover, J., Portillo-Wightman, G., Yeh, L., Havaladar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., and Dehghani, M. (2020). Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 0(0):0.
- Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., and Weber, R. (2021). The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53:232–246.
- Hosseini, S. A., Modgil, S., and Rodrigues, O. (2014). Enthymeme construction in dialogues using shared knowledge. *COMMA*, 14:325–332.
- Hou, Y. and Jochim, C. (2017). Argument Relation Classification Using a Joint Inference Model. In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66, Copenhagen, Denmark. Association for Computational Linguistics.
- Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04*, page 168–177, New York, NY, USA. Association for Computing Machinery.

- Hulpuş, I., Kobbe, J., Meilicke, C., Stuckenschmidt, H., Becker, M., Opitz, J., Nastase, V., and Frank, A. (2019). Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ ISWC*, pages 62–77.
- Hulpuş, I., Kobbe, J., Stuckenschmidt, H., and Hirst, G. (2020). Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online). Association for Computational Linguistics.
- Hulpuş, I., Prangnawarat, N., and Hayes, C. (2015). Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In *International Semantic Web Conference*, pages 442–457. Springer.
- Iyer, R., Koleva, S., Graham, J., Ditto, P., and Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, 7(8):e42366.
- Johnson, K. and Goldwasser, D. (2018). Classification of Moral Foundations in Microblog Political Discourse. In *The 56th Annual Meeting of the Association for Computational Linguistics, ACL’18*, pages 720–730.
- Kanayama, H. and Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 355–363.
- Khoo, C. S. and Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Kobbe, J., Hulpuş, I., and Stuckenschmidt, H. (2020a). Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.
- Kobbe, J., Hulpuş, I., and Stuckenschmidt, H. (2023). Effect graph: Effect relation extraction for explanation generation. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 116–127, Toronto, Canada. Association for Computational Linguistics.
- Kobbe, J., Opitz, J., Becker, M., Hulpuş, I., Stuckenschmidt, H., and Frank, A. (2019). Exploiting Background Knowledge for Argumentative Relation Classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASICs)*, pages 8:1–8:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

- Kobbe, J., Rehbein, I., Hulpuş, I., and Stuckenschmidt, H. (2020b). Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.
- Konjengbam, A., Ghosh, S., Kumar, N., and Singh, M. (2018). Debate stance classification using word embeddings. In Ordonez, C. and Bellatreche, L., editors, *Big Data Analytics and Knowledge Discovery*, pages 382–395, Cham. Springer International Publishing.
- Krippendorff, K. (2011). Computing krippendorff’s alpha-reliability.
- Küçük, D. and Can, F. (2020). Stance detection: A survey. 53(1).
- Lakoff, G. (1997). *Moral Politics: What Conservatives Know That Liberals Don’t*. University of Chicago Press.
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., and Bizer, C. (2015). Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Lewis, P. G. (2019). Moral Foundations in the 2015-16 U.S. Presidential Primary Debates: The Positive and Negative Moral Vocabulary of Partisan Elites. *Social Sciences*, 8(233).
- Li, Y., Garg, K., and Caragea, C. (2023). A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085, Toronto, Canada. Association for Computational Linguistics.
- Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*, 3:111–132.
- Lin, Y., Hoover, J., Portillo-Wightman, G., Park, C., Dehghani, M., and Ji, H. (2018). Acquiring Background Knowledge to Improve Moral Value Prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559.
- Lippi, M. and Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd*

- Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.
- Martinez-Rodriguez, J. L., Lopez-Arevalo, I., and Rios-Alvarado, A. B. (2018). OpenIE-based approach for knowledge graph construction from text. *Expert Systems with Applications*, 113:339–355.
- Matsuo, A., Sasahara, K., Taguchi, Y., and Karasawa, M. (2018). Development of the Japanese Moral Foundations Dictionary: Procedures and Applications. *CoRR*, abs/1804.00871.
- Mayer, T., Cabrio, E., and Villata, S. (2020). Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.
- Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.
- Menini, S. and Tonelli, S. (2016). Agreement and Disagreement: Comparison of Points of View in the Political Domain. In *COLING*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19:1–22.
- Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- More, P. and Ghotkar, A. (2016). A study of different approaches to aspect-based opinion mining. *International Journal of Computer Applications*, 145(6):11–15.
- Murakami, A. and Raymond, R. (2010). Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Coling 2010: Posters*, pages 869–875. Coling 2010 Organizing Committee.

- Nasukawa, T. and Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In *Proceedings of the 2Nd International Conference on Knowledge Capture, K-CAP '03*, pages 70–77, New York, NY, USA. ACM. event-place: Sanibel Island, FL, USA.
- Navigli, R. (2009). Word sense disambiguation. *ACM Computing Surveys*, 41(2):1–69.
- Nguyen, H. and Litman, D. (2016). Context-aware Argumentative Relation Mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany. Association for Computational Linguistics.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Paglieri, F. and Woods, J. (2011). Enthymemes: From reconstruction to understanding. *Argumentation*, 25:127–139.
- Palau, R. M. and Moens, M.-F. (2009). Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL '09*, pages 98–107, New York, NY, USA. ACM.
- Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.
- Paul, D., Opitz, J., Becker, M., Kobbe, J., Hirst, G., and Frank, A. (2020). Argumentative relation classification with background knowledge. In *Computational Models of Argument*, pages 319–330. IOS Press.
- Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.
- Peldszus, A. and Stede, M. (2015). Joint prediction in MST-style discourse parsing for argumentation mining. In *EMNLP*.
- Peldszus, A. and Stede, M. (2016). An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–815, London. College Publications.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Persing, I. and Ng, V. (2016). End-to-End Argumentation Mining in Student Essays. In *HLT-NAACL*.

- Qiu, G., Liu, B., Bu, J., and Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In *Twenty-First International Joint Conference on Artificial Intelligence*. Citeseer.
- Rajendran, P., Bollegala, D., and Parsons, S. (2016). Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39, Berlin, Germany. Association for Computational Linguistics.
- Randolph, J. J. (2005). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss’ fixed-marginal multirater kappa. *Online submission*.
- Rashkin, H., Singh, S., and Choi, Y. (2016). Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321. Association for Computational Linguistics. event-place: Berlin, Germany.
- Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46.
- Razuvayevskaya, O. and Teufel, S. (2017). Finding enthymemes in real-world texts: A feasibility study. *Argument & computation*, 8(2):113–129.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Reisert, P., Inoue, N., Kuribayashi, T., and Inui, K. (2018). Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.
- Rescher, N. (1964). *Introduction to Logic*. St. Martin’s Press.
- Rezapour, R., Shah, S. H., and Diesner, J. (2019). Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 35–45.
- Roggeband, C. and Vliegthart, R. (2007). Divergent framing: The public debate on migration in the Dutch parliament and media, 1995–2004. *West European Politics*, 3(30):524–548.
- Schiller, B., Daxenberger, J., and Gurevych, I. (2020). Stance detection benchmark: How robust is your stance detection?

- Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.
- Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.
- Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50(4):19–45.
- Schwartz, S. H. and Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality*, 38(3):230–255.
- Scriven, M. (1976). *Reasoning*. McGraw-Hill Book Company.
- Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.
- Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.
- Sim, J. and Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3):257–268.
- Sobhani, P., Mohammad, S., and Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 159–169, Berlin, Germany. Association for Computational Linguistics.
- Soderland, S., Roof, B., Qin, B., Xu, S., Mausam, and Etzioni, O. (2010). Adapting open information extraction to domain-specific relations. *AI Magazine*, 31(3):93–102.
- Somasundaran, S. and Wiebe, J. (2009). Recognizing Stances in Online Debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 226–234, Stroudsburg, PA, USA. Association for Computational Linguistics. event-place: Suntec, Singapore.
- Somasundaran, S. and Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

- Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1):4444–4451.
- Sridhar, D., Getoor, L., and Walker, M. (2014). Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117. Association for Computational Linguistics.
- Stab, C. and Gurevych, I. (2014). Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.
- Stab, C. and Gurevych, I. (2017a). Parsing Argumentation Structures in Persuasive Essays. *Comput. Linguist.*, 43(3):619–659.
- Stab, C. and Gurevych, I. (2017b). Recognizing Insufficiently Supported Arguments in Argumentative Essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.
- Stanovsky, G., Michael, J., Zettlemoyer, L., and Dagan, I. (2018). Supervised open information extraction. 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 885–895, New Orleans, Louisiana. Association for Computational Linguistics.
- Stede, M. and Schneider, J. (2018). Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11:1–191.
- Sun, Q., Wang, Z., Zhu, Q., and Zhou, G. (2018). Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Takikawa, H. and Sakamoto, T. (2017). Moral Foundations of Political Discourse: Comparative Analysis of the Speech Records of the US Congress and the Japanese Diet. *CoRR*, abs/1704.06903.
- Tan, L. (2014). Pywsd: Python Implementations of Word Sense Disambiguation (WSD) Technologies [software]. <https://github.com/alvations/pywsd>.
- Thet, T. T., Na, J.-C., and Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848.

- Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335. Association for Computational Linguistics.
- Toledo-Ronen, O., Bar-Haim, R., Halfon, A., Jochim, C., Menczel, A., Aharonov, R., and Slonim, N. (2018). Learning sentiment composition from sentiment lexicons. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2230–2241, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press, 2 edition.
- Van Eemeren, F. H. and Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press.
- Voelkel, J. G. and Feinberg, M. (2018). Morally reframed arguments can affect support for political candidates. *Social Psychological and Personality Science*, 9(8):917–924. PMID: 30595808.
- Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., and Stein, B. (2017a). Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.
- Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017b). 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, Valencia, Spain. Association for Computational Linguistics.
- Walton, D. (1999). Historical origins of argumentum ad consequentiam. *Argumentation*, 13(3):251–264.
- Walton, D. and Macagno, F. (2015). A classification system for argumentation schemes. *Argument & Computation*, 6(3):219–245.
- Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.
- Walton, D. and Reed, C. A. (2005). Argumentation schemes and enthymemes. *Synthese*, 145:339–370.
- Wang, R., Zhou, D., Jiang, M., Jiasheng, S., and Yang, Y. (2019). A survey on opinion mining: from stance to product aspect. PP:1–1.

- Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics. event-place: Vancouver, British Columbia, Canada.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Wood, I., McCrae, J. P., Andryushechkin, V., and Buitelaar, P. (2018). A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Xie, J. Y., Ferreira Pinto Junior, R., Hirst, G., and Xu, Y. (2019). Text-based inference of moral sentiment change. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4654–4663, Hong Kong, China. Association for Computational Linguistics.
- Yang, J., Han, S. C., and Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.
- Yuan, J., Wei, Z., Zhao, D., Zhang, Q., and Jiang, C. (2021). Leveraging argumentation knowledge graph for interactive argument pair identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2310–2319. Association for Computational Linguistics.
- Zapko-Willmes, A., Schwartz, S. H., Richter, J., and Kandler, C. (2021). Basic value orientations and moral foundations: Convergent or discriminant constructs? *Journal of Research in Personality*, 92:104099.
- Özge Sevgili, Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Appendix A

Published Resources

Ch.	Name	Type	Link
3	Debatepedia	Crawled Data	https://madata.bib.uni-mannheim.de/324/
4	StArCon	Crowd-Annotated Data & Code	https://github.com/dwslab/StArCon
5	Effect graph	KG & Crowd- & Expert-Annotated Data	https://github.com/dwslab/Effect-Graph
6	MFs in arguments	Expert-Annotated Data	https://github.com/dwslab/Morality-in-Arguments
6	MFD to WordNet	Expert-Annotated Data	https://github.com/dwslab/Morality-in-Knowledge-Graphs

Overview of resources published with this thesis.

Appendix B

Relevant Dependency Relations

Label	Relation	Description
nsubj	Nominal Subject	“a noun phrase which is the syntactic subject of a clause”
nsubjpass	Passive Nominal Subject	“a noun phrase which is the syntactic subject of a passive clause”
csubj	Clausal Subject	“a clausal syntactic subject of a clause, i.e., the subject is itself a clause.”
csubjpass	Clausal Passive Subject	“a clausal syntactic subject of a passive clause”
dobj	Direct Object	“The direct object of a verb phrase is the noun phrase which is the (accusative) object of the verb.”
cobj	Clausal Object	<i>No description available.</i>
nmod	Nominal Modifier	<i>No description available.</i>
xcomp	Open Clausal Complement	“An open clausal complement (<i>xcomp</i>) of a verb or an adjective is a predicative or clausal complement without its own subject.”
amod	Adjectival Modifier	“An adjectival modifier of an [<i>sic</i>] NP is any adjectival phrase that serves to modify the meaning of the NP.”
nn	Noun Compound Modifier	“A noun compound modifier of an [<i>sic</i>] NP is any noun that serves to modify the head noun.”
advmod	Adverbial Clause Modifier	“An adverbial clause modifier of a verb phrase or S is a clause modifying the verb”.
prep	Prepositional Modifier	“A prepositional modifier of a verb, adjective, or noun is any prepositional phrase that serves to modify the meaning of the verb, adjective, noun, or even another preposition [<i>sic</i>].”
pobj	Object of a Preposition	“the head of a NP following the preposition, or the adverbs ‘here’ and ‘there’”
neg	Negation Modifier	“the relation between a negation word and the word it modifies”

The descriptions are cited from De Marneffe and Manning (2008).

Appendix C

Effect Relation Validity Annotation

The left column presents the arguments from which the effect relations were extracted. The right columns contain the effect relations in subject-verb-object order.

Invalid			
It is good for labeling to turn consumers off to GM foods.	labeling	turn	consumers
Holocaust denial opens doors to harmful, non-factual views	Holocaust denial	opens	doors
Carbon trading wrongly turns atmosphere into tradeable property	Carbon trading	turns	atmosphere
Solar panels are hard to move when a person moves homes.	person	moves	homes
A handgun ban deprives citizens of the most commonly used weapon for self defense	handgun ban	deprives	citizens
Legalization will increase prostitution and subsequently worsen public health.	Legalization	increase	prostitution
Arabs seek pre-1967 borders to weaken and dissolve Israel.	Arabs	seek	pre-1967 borders
Tunnel will increase traffic in pioneer square.	Tunnel	increase	traffic
War in Iraq kept Saddam from acquiring nuclear weapons.	War in Iraq	kept	Saddam
Sequestered CO2 can be injected into reservoirs to recover oil	reservoirs	recover	oil
Somewhat valid			
Drug dealers sell drugs near needle exchanges	Drug dealers	sell	drugs
ICC causes tyrants to cling to power to avoid prosecution	ICC	causes	tyrants
Pageants teach kids to follow rules and play fair.	Pageants	teach	kids
Polygamy reduces the impulse to adultery and resulting divorces	Polygamy	reduces	impulse
Polygamy provides wives with a sisterhood of life-long friends	Polygamy	provides	wives

With pre-1967 borders, PLO would recognize Israel, end conflict	PLO	recognize	Israel
Valid			
Public insurance option will increase taxes, drag-down economy	Public insurance option	increase	taxes
Mars mission would inspire kids to become scientists	Mars mission	inspire	kids
Holocaust denial psychologically harms Holocaust survivors	Holocaust denial	harms	Holocaust survivors
Fear will cause nuclear proliferation, despite testing ban.	Fear	cause	nuclear proliferation
Merit pay gives teachers an incentive to work harder	Merit pay	gives	incentive
Vouchers drain talent from public schools, undermining competitiveness.	Vouchers	drain	talent
The Fairness Doctrine improves the public discourse	Fairness Doctrine	improves	public discourse
Fuel economy standards reduce emissions, fight global warming	Fuel economy standards	reduce	emissions
Global warming grows phytoplankton; growing more is reckless	Global warming	grows	phytoplankton
Global warming kills algae, worsens warming; iron fertilization helps	Global warming	kills	algae
Hybrids significantly reduce emissions, fight global warming	Hybrids	reduce	emissions
Hybrids increase efficiency by shutting engines down while idling.	Hybrids	increase	efficiency
Hybrid vehicles reduce noise pollution	Hybrid vehicles	reduce	noise pollution
Natural gas vehicles reduce emissions, fight global warming	Natural gas vehicles	reduce	emissions
Dams can destroy marine fisheries	Dams	destroy	marine fisheries
NATO expansion threatens and antagonizes Russia	NATO expansion	threatens	Russia
EU enlargement will improve foreign direct investment into eastern Europe.	EU enlargement	improve	foreign direct investment
Prisoner voting undermines punishment and so rehabilitation.	Prisoner voting	undermines	punishment
Space exploration is inspiring and pushes humans to advance	Space exploration	inspiring	humans
Assassinations erode norms against assassination; jeopardizes leaders.	Assassinations	erode	norms
Assassinations protect publics from terrorism; even while it's hard to measure	Assassinations	protect	publics
Abortions encourage infanticide	Abortions	encourage	infanticide
Legal abortion protects women with serious illnesses that are vulnerable.	Legal abortion	protects	women
Animal testing has significantly improved human welfare	Animal testing	improved	human welfare
Sanctions hurt Cuban-Americans with relatives in Cuba.	Sanctions	hurt	Cuban-Americans

By worsening HIV/AIDS, prostitution will devastate societies	prostitution	devastate	societies
Nuclear weapons may protect humans from threats from space.	Nuclear weapons	protect	humans
Ecotourism can damage habitats	Ecotourism	damage	habitats
Ecotourism cultivates a conservation ethic	Ecotourism	cultivates	conservation ethic
Fish feel pain and should not be made to suffer	Fish	feel	pain
High-speed rail allows people to see/visit new places.	High-speed rail	allows	people
Unlike automobiles, rail fosters a sense of community.	rail	fosters	sense of community
National service promotes patriotism.	National service	promotes	patriotism
Mandatory voting would reduce polarization.	Mandatory voting	reduce	polarization
DREAM Act offers citizenship to youth already Americans.	DREAM Act	offers	citizenship
Russia will build-up nuclear arms without New START.	Russia	build	nuclear arms
Law school teaches people to think like lawyers.	Law school	teaches	people
Landmines can protect peacekeepers.	Landmines	protect	peacekeepers
Landmines kill soldiers and limit mobility of military planting them	Landmines	kill	soldiers
Net Neutrality may restrict value-added services	Net Neutrality	restrict	value-added services
Corporate personhood favors corporate interests	Corporate personhood	favors	corporate interests
Corporate personhood enables multinational corporations, global stability.	Corporate personhood	enables	multinational corporations
Corporate personhood protects businesses from discrimination	Corporate personhood	protects	businesses
Small government encourages self-reliance	Small government	encourages	self-reliance
Full-body scans more effectively reveal concealed weapons.	Full-body scans	reveal	concealed weapons
Affirmative action promotes mediocrity by undermining meritocracy.	Affirmative action	promotes	mediocrity

Appendix D

Explanation Quality Annotation

The relations to be explained are written in **bold**. Negative relations are written in *italic*. The validity is encoded as follows: 0 = unvalid, 1 = somewhat valid, 2 = valid. The explanations are ordered descending according their importance (see equation 5.3).

Drug dealers	sell	drugs			Importance	Validity
drug dealer	add	substance	hyponym	drug	2.77E-03	0
drug dealer	sell	weed	hypernym	drug	1.11E-03	1
drug dealer	grow	marijuana	hypernym	drug	9.05E-04	0
drug dealer	sell	marijuana	hypernym	drug	9.05E-04	1
drug dealer	hypernym	somebody	lend	drug	1.18E-04	0
drug dealer	hypernym	somebody	sell	drug	1.18E-04	0
drug dealer	hypernym	person	feel	drug	1.04E-04	0
drug dealer	hypernym	person	like	drug	1.04E-04	0
drug dealer	hypernym	person	take	drug	1.04E-04	0
drug dealer	hypernym	individual	take	drug	8.98E-05	0
drug dealer	hypernym	someone	administer	drug	7.08E-05	0
drug dealer	hypernym	someone	bring	drug	7.08E-05	0
drug dealer	hypernym	someone	buy	drug	7.08E-05	0
drug dealer	hypernym	someone	give	drug	7.08E-05	0
drug dealer	hypernym	someone	offer	drug	7.08E-05	0
drug dealer	hypernym	someone	put	drug	7.08E-05	0
drug dealer	hypernym	someone	slip	drug	7.08E-05	0
drug dealer	hypernym	someone	take	drug	7.08E-05	0
drug dealer	hypernym	criminal	obtain	drug	4.82E-05	0
Pageants	teach	kids				
pageant	take	part	hyponym	kid	4.69E-06	0
pageant	encourage	child	synonym	kid	6.08E-07	0
Space exploration	inspiring	humans				
space exploration	benefit	humanity	show	human	5.08E-04	0
space exploration	improve	technology	allow	human	2.81E-04	2
space exploration	improve	technology	help	human	2.81E-04	2
space exploration	improve	technology	save	human	2.81E-04	2
space exploration	benefit	humanity	attribute	human	6.01E-05	1
space exploration	benefit	humanity	hypernym	human	6.01E-05	1
space exploration	benefit	humanity	nominalization	human	6.01E-05	1
space exploration	allow	man	hypernym	human	3.51E-06	0
space exploration	allow	man	nominalization	human	3.51E-06	0
space exploration	allow	man	synonym	human	3.51E-06	0
space exploration	inspire	child	develop	human	1.79E-06	0
Abortions	encourage	infanticide				
abortion	consider	murder	hyponym	infanticide	1.20E-03	0
abortion	fit	murder	hyponym	infanticide	1.20E-03	0
abortion	justify	murder	hyponym	infanticide	1.20E-03	0
Legal abortion	protects	women				
legal abortion	<i>deny</i>	fetus	<i>damage</i>	woman	1.10E-04	2
legal abortion	<i>deny</i>	fetus	<i>endanger</i>	woman	1.10E-04	2

legal abortion	<i>deny</i>	fetus	<i>harm</i>	woman	1.10E-04	2
Polygamy	provides	wives				
polygamy	put	somebody	hyponym	wife	5.67E-04	0
polygamy	cause	person	hyponym	wife	2.56E-05	0
polygamy	influence	person	hyponym	wife	2.56E-05	0
polygamy	allow	man	get	wife	2.14E-05	0
polygamy	allow	man	marry	wife	2.14E-05	0
polygamy	allow	man	take	wife	2.14E-05	0
polygamy	encourage	man	get	wife	2.14E-05	0
polygamy	encourage	man	marry	wife	2.14E-05	0
polygamy	encourage	man	take	wife	2.14E-05	0
polygamy	cause	woman	hyponym	wife	1.99E-06	0
prostitution	devastate	societies				
prostitution	<i>hurt</i>	traditional marriage	help	society	4.76E-02	2
prostitution	<i>hurt</i>	traditional marriage	make	society	4.76E-02	2
prostitution	pay	prostitute	<i>hurt</i>	society	3.09E-03	0
prostitution	nominalization	prostitute	<i>hurt</i>	society	1.37E-03	0
prostitution	hypernym	act	<i>harm</i>	society	3.31E-04	0
prostitution	hypernym	crime	<i>hurt</i>	society	2.78E-04	2
prostitution	<i>harm</i>	marriage	benefit	society	1.19E-04	2
prostitution	<i>harm</i>	marriage	form	society	1.19E-04	2
prostitution	allow	man	<i>hate</i>	society	7.13E-06	0
Nuclear weapons	protect	humans				
nuclear weapon	protect	human race	hypernym	human	2.50E-03	1
nuclear weapon	make	world	hypernym	human	3.18E-06	0
nuclear weapon	help	world	hypernym	human	3.18E-06	1
PLO	recognize	Israel				
plo	region domain	palestine	recognize	israel	1.60E-01	1
plo	region domain	palestine	take	israel	1.60E-01	0
plo	region domain	palestine	convince	israel	1.60E-01	0
Fish	feel	pain				
fish	hypernym	animal	feel	pain	3.59E-02	2
fish	hypernym	animal	receive	pain	3.59E-02	1
fish	hypernym	organism	feel	pain	4.76E-03	2
fish	<i>lose</i>	ability	<i>fall</i>	pain	2.88E-04	0
fish	hypernym	person	feel	pain	1.37E-05	0
fish	hypernym	someone	accept	pain	3.92E-06	0
fish	hypernym	someone	cause	pain	3.92E-06	0
fish	hypernym	someone	feel	pain	3.92E-06	0
high-speed rail	allows	people				
high-speed rail	offer	freedom	allow	people	4.82E-05	2
high-speed rail	save	life	form	people	5.27E-06	0
Russia	build	nuclear arms				
russia	back	iran	get	nuclear arm	1.98E-04	0
russia	give	iran	get	nuclear arm	1.98E-04	0
russia	provide	iran	get	nuclear arm	1.98E-04	0
russia	reserve	right	produce	nuclear arm	2.78E-06	0
Law school	teaches	people				
law school	hypernym	institution	treat	people	3.09E-03	0
law school	hypernym	group	educate	people	3.95E-04	0
law school	hypernym	school	take	people	2.04E-04	0
law school	hypernym	school	teach	people	2.04E-04	2
law school	hypernym	school	tell	people	2.04E-04	0
law school	hypernym	school	encourage	people	2.04E-04	1
Landmines	kill	soldiers				
landmine	<i>kill</i>	animal	hyponym	soldier	4.83E-06	0