Content Selection for Effective Counter-Argument Generation

Christopher Hidey

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Content Selection for Effective Counter-Argument Generation

Christopher Hidey

The information ecosystem of social media has resulted in an abundance of opinions on political topics and current events. In order to encourage better discussions, it is important to promote high-quality responses and relegate low-quality ones. We thus focus on *automatically* analyzing and generating counter-arguments in response to posts on social media with the goal of providing effective responses.

This thesis is composed of three parts. In the first part, we conduct an *analysis of arguments.* Specifically, we first annotate discussions from Reddit for aspects of arguments and then analyze them for their persuasive impact. Then we present approaches to identify the argumentative structure of these discussions and predict the persuasiveness of an argument. We evaluate each component independently using automatic or manual evaluations and show significant improvement in each.

In the second part, we leverage our discoveries from our analysis in the process of *generating counter-arguments.* We develop two approaches in the retrieve-and-edit framework, where we obtain content using methods created during our analysis of arguments, among others, and then modify the content using techniques from natural language generation. In the first approach, we develop an approach to retrieve counter-arguments by annotating a dataset for stance and building models for stance prediction. Then we use our approaches from our analysis of arguments to extract persuasive argumentative content before modifying non-content phrases for coherence. In contrast, in the second approach we create a dataset and models for modifying content – making semantic edits to a claim to have a contrasting stance. We evaluate our approaches using intrinsic automatic evaluation of our predictive models and an overall human evaluation of our generated output.

Finally, in the third part, we discuss the *semantic challenges of argumentation* that we need to

solve in order to make progress in the understanding of arguments. To clarify, we develop new methods for identifying two types of semantic relations – causality and veracity. For causality, we build a distant-labeled dataset of causal relations using lexical indicators and then we leverage features from those indicators to build predictive models. For veracity, we build new models to retrieve evidence given a claim and predict whether the claim is supported by that evidence. We also develop a new dataset for veracity to illuminate the areas that need progress. We evaluate these approaches using automated and manual techniques and obtain significant improvement over strong baselines. Finally, we apply these techniques to claims in the domain of household electricity consumption, mining claims using our methods for causal relations and then verifying their truthfulness.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my gratitude to my advisor, Kathleen McKeown for her mentorship and guidance. I will always appreciate that she encouraged me to pursue new directions and helped me to learn to conduct research independently. Most importantly, I appreciate that she created a positive work environment in the NLP lab that allowed for different approaches and working habits. I also wish to thank Smaranda Muresan and Mona Diab for their mentorship and the rest of the thesis committee, Owen Rambow and Julia Hirschberg, for their advice and support.

I also appreciate the friends and colleagues who have provided advice and support during my time at Columbia. I especially wish to thank Tuhin Chakrabarty, Alyssa Hwang, Elena Musi, Siddharth Varia, Tariq Alhindi, Kriste Krstovski, and Gengyu Wang, whose contributions made much of the work in this thesis possible. Thanks also to Chris Kedzie, Jessica Ouyang, Noura Farra, Elsbeth Turcan, Fei-Tzin Lee, Emily Allaway, Or Biran, Yves Petinot, Kapil Thadani, Debanjan Ghosh, Olivia Winn, Katy Gero, Giannis Karamanolakis, and Tom Effland for informal discussions, reading groups, paper reviews, and many other things.

Finally, I am grateful for my family. Mom, Dad, Adam, Celeste, Jeremy, and Stephanie, thanks for always being there and love you all. Sharon, thanks for being so supportive during this journey, this would not be possible without you. Audrey and Evelyn, thanks for being a continuously-growing reminder of how long I have been a PhD student. I hope you can read this someday.

Chapter 1: Introduction and Background

For many people, social media is their primary source of information, making it a key venue for opinionated discussion [Matsa and Shearer, 2018]. Recent research has shown that 62% of American adults get their news from social media [David et al., 2019], resulting in a population that looks to the arguments of others to determine what they should do and what they should think about current events and topics. The reliance on social media as a source has provided an ecosystem for opinions with a number of different challenges. For example, as of March 2020, the rise of the novel coronavirus COVID-19 has necessitated an interest in what beliefs one should have and actions one should take in response. Providing opinions in this area is important not just in stating *what* to do or think but also *why*. Furthermore, elevating high-quality arguments is key to helping people understand the impact of the virus and their role in minimizing its impact. Second, discussions about politics are especially relevant during an election year. Providing counter-arguments against each side of an issue or a policy is necessary for voters to make educated decisions. Finally, beliefs about topics such as vaccines may result in detrimental effects on societal behavior. Vaccines are a highly controversial topic, in spite of overwhelming evidence for their benefits, and proponents of avoiding vaccination may rely on misinformation and other techniques to persuade others.

In order to address the challenges discussed in these examples, it becomes important to provide effective responses in online debates. Addressing the first example requires identifying opinions and their supporting evidence, as well as discerning what makes some arguments more effective than others. Addressing the issue from the second example requires the ability to provide a range of *counter-arguments*, as many people have stated they prefer multiple points of view rather than a single-viewpoint source such as a newspaper [Newman et al., 2014]. Furthermore, the way that information is presented is key [Brossard, 2013] – uncivil responses may cause negative perception towards a topic [Gerhards and Schäfer, 2010, Runge et al., 2013]. Finally, in the domain of the third example, the identification of accurate information is increasingly important to counter the rise of misinformation [Ireton, 2018].

Our goals, then, are to be able to automatically provide accurate, novel information in support of an opinion in an effective way. To make progress towards this goal, this thesis introduces contributions in a number of different areas. First, we make contributions in the *analysis of arguments*, providing answers to questions such as what makes an argument persuasive and how to identify arguments and their supporting evidence. Second, we introduce methods for the *generation of counter-arguments*, incorporating the results of our analysis of arguments into our approaches. Third, we discuss the challenges of the *semantics of arguments* along two dimensions – causality and veracity – and introduce methods for identification of these types of relations along with an application of our methods.

In Section 1.1, we discuss the background and terminology needed for argumentation and persuasion and provide a comparison to other types of communication. We also introduce our end task of counter-argument generation. Section 1.2 discusses ethical issues under consideration when dealing with opinions in online debates. Lastly, in Section 1.3, we discuss the contributions of the thesis and the organization of individual sections.

1.1 Background and Terminology

Argumentation is a type of discourse where speakers try to persuade their audience about the reasonableness of a claim by producing supporting arguments [van Rees, 2007]. From a theoretical perspective, researchers have proposed different views of argumentation. The key components of the Toulmin [1958] model are the claim, backing, and (implicit) warrant, with rebuttal as an additional necessary component in a dialogue. This underlying structure determines the form of the argument while the topic and other claims and premises determines the content. On the other hand, Freeman [1991] views an argument as an exchange between an proponent and an opponent, where the proponent presents claims and defends them with reasons and the opponent attacks them.

While the terminology differs across these models, these approaches view an argument as a series of reasons leading towards a conclusion. In this thesis, we follow other work in computational argumentation and primarily use the Toulmin [1958] model, referring to premises that support claims and in turn support a main claim or conclusion. However, we also incorporate the view of Freeman [1991] as our work addresses social media dialogues.

We additionally focus on persuasion, specifically in online debates. Research shows that individuals do in fact update their views in the direction of information (i.e. they are Bayesians) [Coppock, 2016], making the identification and presentation of novel information key to a successful argument. While those with strong beliefs may update their views in small increments [Coppock, 2016], our work assumes that participants in a debate are already open to changing their views. Even though some argumentation online is in bad faith (e.g. ad hominems [Habernal et al., 2018b]), we focus only on good-faith, effective arguments found in moderated debate forums where the goal is to persuade.

Argumentation is distinct from persuasion as the latter may occur without a well-constructed sequence of reasoning. As underlined in Rhetorics and Argumentation Theory [Perelman and Olbrechts-Tyteca, 1973, van Eemeren and Eemeren, 2009], the persuasiveness of a message lies at the interface between discourse form (e.g., use of hedges, connectives, rhetorical questions, and structure) and conceptual form such as the artful use of ethos (credibility and trustworthiness of the speaker), pathos (appeal to audience feelings), and logos (appeal to the rationality of the audience through logical reasoning). While these aspects are present in monologic argumentation (e.g. persuasive essays), which has been the object of extensive focus in computational argumentation research, dialogues have additional challenges. Both monologues and dialogues require an understanding of intra-argument interaction, but the forms of dialogue include inter-argument interaction as well.

In order to understand and validate the effectiveness of arguments, then, we build models of both form and content using the semantics and pragmatics of arguments. Effective argumentation is more than just form and content - the role of style [Wang et al., 2017], framing, [Hartmann et al., 2019] agreement [Hidey et al., 2017], and other factors are key - but we need to be able to study form and content without being overly concerned about differences in style and domain. The focus of this thesis, then, is not on generation but on content retrieval and identification.

As a step towards counter-argument generation, our approach follows a *retrieve-and-edit* approach to argumentation, which involves obtaining and modifying existing content from the Web. We provide methods for identifying persuasive counter-arguments and develop approaches to extract key content from these arguments by identifying claims, premises, and persuasive argumentative components. We also develop approaches to obtaining content through semantic relations – causality and veracity. We then apply aspects of these approaches towards counter-argument generation. We validate the effectiveness of arguments through both intrinsic and extrinsic evaluations. For intrinsic evaluations, we evaluate our models of form and content on datasets labeled for aspects of argumentation or semantic relations. For extrinsic evaluations, we use automated metrics and human evaluations of generated or retrieved arguments.

We explore several questions about the effectiveness of counter-arguments by examining the role of content:

- How are arguments structured and what makes an effective counter-argument? We evaluate via crowd-sourcing and automated metrics.
- How can we retrieve and generate counter-arguments? We incorporate the analysis of arguments along with models for contrastive meaning.
- What role do semantic relations play in argumentation? We answer this question via intrinsic evaluation on respective datasets and an extrinsic application.

Consider the example in Figure 1.1 from the online discussion forum "Change My View". In this discussion, the original poster (OP) states his/her belief (claim) that bicyclists are subsidized by car owners for two reasons (backing or premises): car owners pay road taxes and bike lanes take space away from cars. The responder (R) provides sources with data on the true cost to the taxpayer of drivers, cyclists, and others (a rebuttal). Finally, the OP acknowledges that R has made convincing points.

OP makes a claim and provides non-argumentative background. They support their claim with logical reasoning, although they make additional claims with emotional connotations as well. They then link these propositions[1] together using argumentative "shells" [Madnani et al., 2012] such as discourse connectives. When R generates a response, R retrieves evidence from external sources that contradict the claims of OP, rather than addressing all of the premises. They conclude their argument by agreeing with OP in part, but by making a contrastive claim that disagrees with the main claim of OP.

Consider the sub-tasks necessary to generate a counter-argument to OP. First, R identifies the **claims** of OP and the supporting **premises**. OP's post contains a significant amount of non-argumentative content (background on the conversation with the Uber driver and other asides) and R needs to determine the argumentative content to respond accordingly. Second, this rebuttal requires an understanding of contrastive meaning, both in terms of retrieving evidence with a different **stance** and providing **counter-claims**. Rather than addressing all the points of OP, R finds evidence that directly contradicts the main claim of OP and states the opposite view. Next, R needs to verify the evidence is true by **fact-checking** the household cost paid to cover drivers and needs to logically connect the sequence of reasoning about car-share-companies using the explicit connectives *"because"* and *"for two reasons,"* which represent **causal relations**. Finally, R **fuses** their multiple arguments together by using shell phrases, explicit discourse connectives like *"So"* and *''Ultimately,"* along with speech acts such as agreement (e.g. *"I can't find any flaws in his argument"* and *"cyclists are being subsidized."*) The author puts all this together in a globally coherent way that is **persuasive** to OP.[2] In order to understand what makes the argument effective, we need to consider the interaction with the original post to predict persuasion as well as the most persuasive claims and premises from the counter-argument.

This list of conditions is necessary (although not sufficient) for understanding argumentation.

[1]We use the terminology of formal logic where a proposition is a statement with a truth value.
[2]as indicated by OP in the Change My View forum

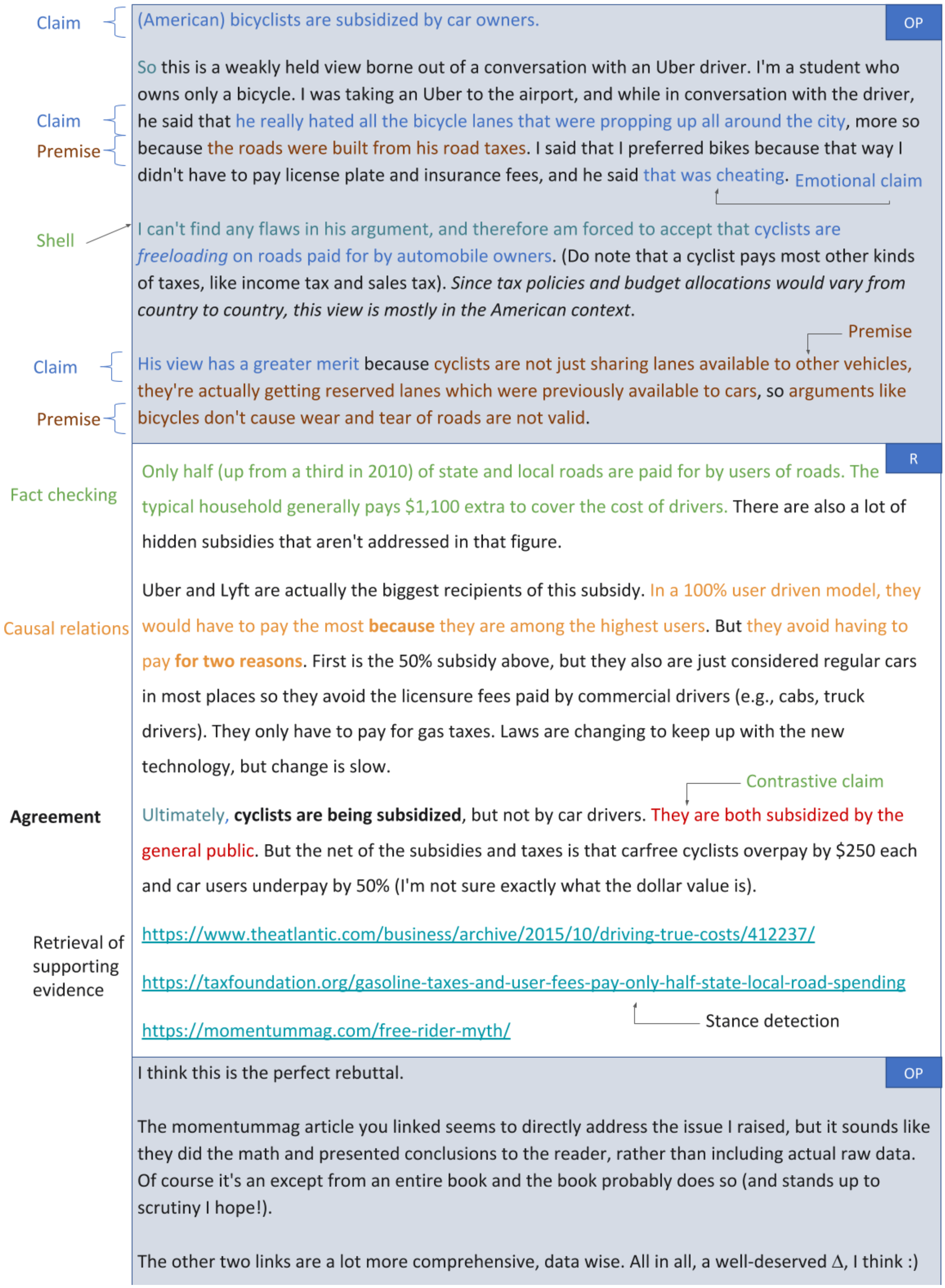


Figure 1.1: Truncated Discussion Thread from Change My View

Other considerations include framing [Hartmann et al., 2019] – this discussion is framed as an issue of fairness, determining whether drivers and cyclists are paying their fair share, although one alternative would be as a matter of logistics. Additionally, other types of discourse semantics are shown in this example – R uses temporal reasoning (*"up from a third in 2010"*) and provides examples (Uber and Lyft). Personal narrative plays a role as well – OP relates a story but not for a persuasive intent. There are also many rhetorical devices not shown here, such as the use of rhetorical questions or analogies. Finally, presentation issues such as style (e.g. formality and hedging) may affect how the argument is communicated and received. The effective presentation of the argument may be dependent on the prior beliefs [Durmus and Cardie, 2018a] of OP as well as their personality type [Lukin et al., 2017] – given this information, emotional or logical responses may be more effective. However, we do not consider these aspects in our work.

The provided example requires many aspects of argumentation that would be difficult to encode in a single computational model. Hence we focus on a retrieve-and-edit approach where we obtain candidate arguments using Google or another search engine. This approach allows us to obtain content, to which we can then apply predictive models such as those for stance detection or persuasiveness to identify and extract arguments, and then modify them as needed to present coherently as counter-arguments. Thus we analyze arguments in order to understand what makes an effective argument so that we can generate a response.

1.2 Ethical Considerations

Our approaches to argument analysis and generation must be conducted with a view towards ethics. Research in natural language processing has been concerned with the use or misuse of data, in terms of bias and fairness [Zhao et al., 2017]. While we do not directly study these issues, the focus on analysis and retrieval of arguments allows one to incorporate approaches for mitigating these issues in future work. Instead, we primarily try to avoid embodying the negative behaviors of online users in our models. We consider three issues that should be accounted for in computational argumentation: 1) experiments should distinguish persuasion from propaganda 2) arguments should not contain abusive language and 3) models should be as transparent as possible.

While there may be concerns about building computational models of persuasion, social scientists usually distinguish persuasion from its more nefarious counterparts such as propaganda and manipulation. First, Nettel and Roque [2012] distinguishes persuasive argumentation from manipulation based on two qualities: dissimulation and constraint. Dissimulation, the concealment of intention, involves the hiding of information. Constraint, on the other hand, involves removing the options of the recipient of an argument so that they have no choice but to accept the conclusion. In contrast, we are concerned with *providing novel information* about a topic, rather than limiting information or forcing a desired action. Second, Jowett and O'Donnell [1986] define propaganda as a form of communication that only benefits the propagandist. While propaganda furthers the intent of the propagandist, persuasion "promotes mutual fulfillment" for both the sender and recipient of information. Likewise, in our work, we attempt to benefit the recipient by making them more informed about a topic.

Recent research in natural language processing has been focused on the identification of abusive language in social media [Chakrabarty et al., 2019a]. While there has been significant progress in the identification of abusive language, preventing neural models from generating harmful statements is a difficult problem. For example, Microsoft created a chatbot named Tay which quickly learned from internet dialogues to generate racist and sexist language.[3] Additionally, recent work has shown that models can be easily guided to generate abusive content [Wallace et al., 2019]. The need to minimize these issues provides additional justification for our retrieve-and-edit approach to argument generation. As generative models are often not controllable in terms of their content, retrieval-based approaches allow for the identification of problematic content that can be discarded as one step in the full pipeline.

Finally, transparency is key for building trust in computational models. Due to improvements in dialogue modeling, it is increasingly difficult for Internet users to distinguish humans from robots.[4] Consequently, California recently passed the "B.O.T" (Bolstering Online Transparency)

[3] `https://en.wikipedia.org/wiki/Tay(bot)`
[4] `https://www.theguardian.com/technology/2018/nov/18/how-can-you-tell-who-is-`

law, which requires a chatbot to identify itself when deployed online.[5] Due to these concerns, some researchers have raised questions whether large pre-trained models should be released to the public [Radford et al., 2019]. However, other research has shown that these models are also the best defense against malicious adversarial chatbots for fake news [Zellers et al., 2019]. Thus, it is important to understand what makes an argument persuasive and how to generate arguments in order to counter propaganda and misinformation. This process is analogous to that of cyber-security, where defenders are trying to build systems and adversaries are trying to break them by exploiting weaknesses [Ruef et al., 2016].

1.3 Contributions and Organization of Thesis

Given the aforementioned goals and limitations of the analysis and generation of arguments, we provide contributions in three main areas: 1) the **analysis of arguments** 2) the **generation of counter-arguments** and 3) identifying the **semantic challenges** of arguments.

We create a dataset and models for the **analysis of arguments** towards the goal of providing effective counter-arguments. We annotate data in an online discussion forum for argumentative structure and build models to identify this structure. We also leverage data labeled for persuasive impact to analyze the semantics of arguments and build models for predicting the persuasive effect of arguments.

Our approach to the **generation of counter-arguments**, then, relies on our analysis of argument structure and persuasion. We take two complementary approaches to counter-arguments, both in a retrieve-and-edit setting. For the first, we annotate a dataset for the stance of claim pairs. We then build models for stance detection and use this model in combination with our models of structure and persuasion to extract and edit arguments from multiple sources. For the second approach, rather than retrieving claims with opposite stance, we edit them to have contrastive meaning. This second model could also be used as part of a retrieval-based framework, where we would edit and

human-online-chatbots

[5] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill¡d=201720180SB1001

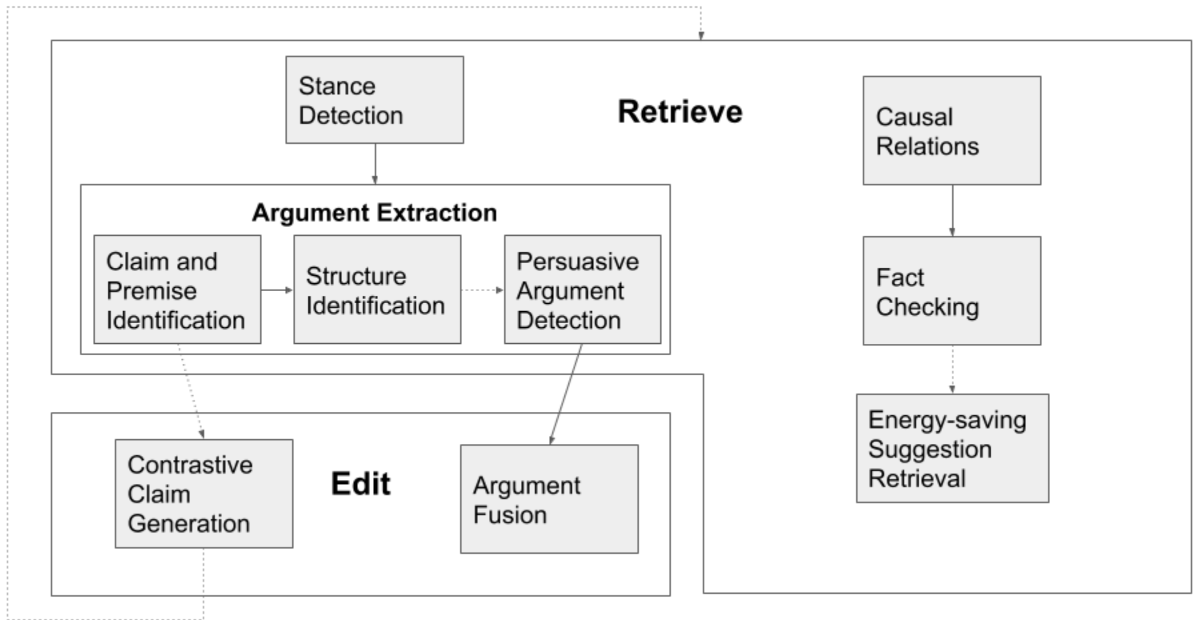


Figure 1.2: Interaction of Different Components of the Thesis

retrieve, rather than the converse for the first model.

Finally, as our approaches to argument generation will demonstrate some of the semantic limitations of current models, we discuss the **semantic challenges for argumentation** that need to be addressed to make progress in the understanding and generation of arguments. For one aspect, causality, we create a dataset and model for identifying causal relations. For the second aspect, veracity, we create an adversarial dataset demonstrating weaknesses in existing approaches to fact-checking and build a model that addresses some of these issues. Finally, we examine an application in the domain of household electricity consumption that requires an understanding of both causality and veracity. We provide a method for the retrieval of claims using our systems developed in each of these areas.

Figure 1.2 diagrams the interaction of these different components. Solid lines indicate that we use a component directly as part of another, whereas dashed lines indicate that this component only influenced our decision-making. In our approach, we assume access to some candidate pool of arguments, both for and against a topic but with unknown stance, which may be retrieved using

Chapter	Description	Section	Contribution	Publication(s)
4	Analysis of Arguments	4.1	Annotation of argument components, semantic types, and relations	Hidey et al. [2017] Chakrabarty et al. [2019c]
		4.2	Prediction of argument components and relations	Chakrabarty et al. [2019b] Chakrabarty et al. [2019c]
		4.3	Prediction of persuasive influence	Hidey and McKeown [2018]
5	Generation of Arguments	5.1	Annotation and prediction of stance	TBD
		5.1	Multi-argument extraction and fusion	TBD
		5.2	Stance modification	Hidey and McKeown [2019]
6	Semantics of Arguments	6.1	Causal relations	[Hidey and McKeown, 2016]
		6.2	Fact-checking	[Hidey and Diab, 2018]
		6.3	Application to household electricity consumption	TBD

Table 1.1: Contributions of the Thesis

a search engine such as Google or another method. Given a set of arguments, we need to identify the stance along with claims and premises. Then, as part of counter-argument generation, if the claim has the same stance we can edit the claim to have a contrastive view. Alternatively, if it has an opposing stance we can extract an entire argument based on its persuasive aspects and edit the argument for discourse connectives and other aspects of presentation. As another option, we can use causal relations to identify claims about a topic and fact-check for veracity.

Table 1.1 outlines the main contributions in each of the three areas along with the corresponding sections and publications. In Chapter 2, we discuss related research in argumentation, generation, and semantics. In Chapter 3, we describe our sources of data for online debates and information as well as specific resources for obtaining this data. Chapters 4, 5, and 6 discuss our approaches to the analysis, generation, and semantics of arguments, respectively. Finally, we discuss the limitations of our approach and future work in Chapter 7.

Chapter 2: Related Work

The predecessors to analyzing arguments computationally include tasks such as opinion mining and sentiment analysis. The objective of opinion mining is to distinguish between opinionated and non-opinionated statements; Liu [2010] defines opinion mining as the computational study of opinions, sentiments, and emotions in text, where an opinion is a “belief or perception about reality.” The task of sentiment analysis, although sometimes used interchangeably, requires the identification of positive and negative views [Lawrence and Reed, 2019]. Argumentation is distinct from sentiment and opinion mining as the latter is largely only concerned with identifying what opinions are being expressed rather than why people hold certain opinions [Lawrence and Reed, 2019] while argumentation theory is concerned with the justification of opinions. Additionally, our work focuses on natural language argumentation rather than formal models of argument [Kowalski and Toni, 1996]. We do follow the terminology of formal logic, referring to *propositions*, or statements with truth values. However, formal models incorporate elements of propositional logic [Rahwan and Larson, 2008], game theory [Rahwan et al., 2009], or uncertainty [Rienstra et al., 2013] for richer representations, while argumentation in language is primarily only concerned with the role of and relation between textual propositions.

As an argument is typically defined as a set of claims, where a subset of those claims – the premises – support another claim – the conclusion [Govier, 2010], a significant body of work has been devoted to uncovering the underlying structure in an argument. In Section 2.1, we discuss overall models of argumentation and how structure often informs quality and effectiveness. We also discuss what makes an argument persuasive and distinguish persuasion from other aspects of online discussions such as conversation quality, influence, or power. We then describe previous work in the annotation of argumentative data for modeling and the common domains where argumentation appears, the challenges associated with that process, and alternatives to manual annotation. We next discuss previous work in computational models of argument and the aspects needed to properly characterize an argument, including structure (Section 2.1.2), propositional content (2.1.3), and realization (2.1.4).

In Section 2.2, we cover related work in generation, discussing recent advances in modeling text-to-text generation using neural networks, as well as the challenges associated with creative text generation, where the input and output are not highly correlated (which is the case in argument generation), unlike a task such as machine translation. We also discuss recent work specifically in argument generation, which includes elements of retrieval (Section 2.2.1), discourse and dialogue planning (Section 2.2.2), and realization (Section 2.2.3).

Finally, in Section 2.3 we discuss the role of semantic relations in argumentation, with a focus on contrastive meaning (Section 2.3.1) and causality (Section 2.3.2), and the challenges that need to be addressed in order to make progress towards full natural language understanding of arguments.

2.1 Argumentation

A significant amount of initial work in computational argumentation has focused on identifying the structure of documents: claims and premises and the relations between them. More recently, there has been increased focus on topics such as the overall properties of argumentation (such as quality or persuasiveness) or intrinsic properties of argument content (such as emotion or credibility). We thus discuss overall argumentation in Section 2.1.1. We then describe relevant work in structure (Section 2.1.2), content (Section 2.1.3), and realization (Section 2.1.4). Our work lies at the intersection of the three. The combination of these aspects is necessary for overall argumentation. Finally, we discuss prior approaches to annotating and obtaining data for argumentative tasks (Section 2.1.5).

2.1.1 Overall Argumentation

The aspects of high-quality arguments and those of persuasive arguments are distinct but not disjoint. While an argument may be evaluated based on how well it gives reasons to provide knowledge about a subject [Nettel and Roque, 2012], a well-written argument may not be effective at convincing the audience to change their views. Conversely, a persuasive argument may appeal to the audience using an argumentative fallacy but nevertheless be successful.

Argument Quality Wachsmuth et al. [2017c] introduced holistic measures of argument quality by consolidating previous theoretical work on argument quality. Their approach unites aspects of three areas: logical – elements such as the presence of supporting evidence or logical fallacies, rhetorical – devices including emotion or clarity, and dialectical – consensus-seeking measures like reasonableness. They created a corpus using graduate-level expert annotators and found low agreement for subjective aspects such as emotional effectiveness and high agreement for aspects of overall quality. One of the dialectical dimensions that Wachsmuth et al. [2017c] identify is convincingness, which is distinct from persuasion as it requires group consensus; previous work [Habernal and Gurevych, 2016a] used Mechanical Turk workers to rank Internet arguments for this aspect and in their analysis found features such as word length and sentiment to be useful. Other research [Simpson and Gurevych, 2018] confirmed these results using a Bayesian preference approach. In later work, Habernal and Gurevych [2016b] further annotated the *reasons* provided by the Turkers for convincingness and predicted the labels using neural methods; a further study [Wachsmuth et al., 2017b] on the same dataset found high correlation between expert-annotated dimensions of overall quality and non-expert reasons for convincingness. Gleize et al. [2019] and Toledo et al. [2019] created crowdsourced corpora of convincingness and argument quality (using the platform Figure Eight), respectively, by directly asking annotators whether they would recommend the use of a particular argument and which of two arguments they would prefer, asking them to disregard their own opinion on a topic in the latter case; they then modeled the arguments using neural approaches. In other work, Durmus et al. [2019b] used crowdsourced argument trees[1] for a particular claim and predicted those claims that have the most impact according to audience votes; they found that modeling the context of a claim is necessary. Finally, Wei et al. [2016] learn to rank arguments according to upvotes in Change My View and find that length features

[1] http://kairos.com

(such as the number of words and paragraphs) and argumentative features (such as the cosine similarity between arguments) were the most helpful. This line of research suggests that overall measures of argument quality can be successful but subjective measures of argument effectiveness should consider the point of view of the originator of the argument when possible; our work on personalized persuasion thus also models the source of the argument and the interaction with the response. Furthermore, our work also leverages structural features of the argument (e.g. percentage of unsupported claims) as well as the post (e.g. number, length, and sequencing of posts).

Other work has attempted to predict overall measures of "helpfulness" in peer reviews [Xiong and Litman, 2011] or product reviews [Danescu-Niculescu-Mizil et al., 2009] or the effectiveness of student essays. Deane [2013] raises the question of whether the automated scoring of persuasive essays adequately captures the construct of writing. Ong et al. [2014] demonstrated a correlation between argument diagrams and essay scores, using a rule-based approach to predict both the structure, e.g. by using discourse connectives to identify support/oppose relations, and score, e.g. by awarding points for support/oppose. Wachsmuth et al. [2016] and Ghosh et al. [2016] showed the efficacy of using structural features to predict essay scores given both gold and predicted argumentative components. In other work [Stab and Gurevych, 2016, 2017b], researchers predict one of the logical aspects identified by Wachsmuth et al. [2017b] in persuasive essays - sufficiency, whether an argument is well-supported by evidence. Finally, recent work [Song et al., 2014, Beigman Klebanov et al., 2016, 2017] has progressed toward linking structure with content by using argumentation schemes [Walton, 1995], or patterns of reasoning, and showing that schemes such as causal reasoning and suggesting alternatives are positively correlated with essay scores, while generic sentences are negatively correlated. In our work, we evaluate the interaction between gold argumentative structure and semantic aspects of arguments on persuasive dialogues [Hidey et al., 2017]. In addition, we develop models for predicting when a causal relation is present [Hidey and McKeown, 2016].

Persuasion Argumentation quality is distinct from persuasion in that the goal of identifying the former is to find an objective, communal evaluation of an argument whereas the goal of the latter is to identify changes in beliefs or actions.

One area of research has examined personalized persuasion using self-reported changes in views (as we also do in our work). Tan et al. [2016] used Change My View to predict which of two counter-arguments to a post is persuasive, as indicated by the original poster. They found that interaction features (e.g. overlap in style words and novel content words) and length features (e.g. number of sentences) were indicative of winning arguments. Jo et al. [2018] followed this work by using neural methods to model the weak points of an argument and whether the response addresses them. Our work on Change My View also models the interaction between arguments [Hidey and McKeown, 2018]. Egawa et al. [2019] annotated "elementary units" in Change My View such as fact and testimony but found no difference in strategies between winning and losing arguments. Our work, however, found that the use of modes of persuasion (e.g. emotion and reason) distinguished winning arguments [Hidey et al., 2017].

Another area of work has examined the effect of debates on the views of an audience. Shirafuji et al. [2019] used argument similarity and discourse features to predict the winner of online debates[2] according to which argument changed the most views. Cano-Basave and He [2016] analyzed the effect that political debates had on poll numbers and found that influential speakers talk for longer periods of time and that structural argumentative features improve over a unigram baseline. Other work has used transcripts from Oxford-style debates[3] to predict winning arguments, which are determined according to the change in audience stance before and after the debate. Researchers have modeled these debates using conversation flow – coverage of one's own arguments and their opponent's [Zhang et al., 2016b], the interaction between style and latent content [Wang et al., 2017], and a turn-based "bag of talking points" approach using recurrent neural networks [Potash and Rumshisky, 2017]. We also use recurrent neural models for our persuasiveness prediction models [Hidey and McKeown, 2018].

[2] `http://idebate.org`

[3] `http://www.intelligencesquaredus.org`

Other research has examined traits of the debater and audience such as personality, demographics, and prior beliefs. Lukin et al. [2017] found that the persuasive effect of emotional and factual arguments depend on audience personality traits. El Baff et al. [2018] created a new corpus that models view change as a spectrum - whether news editorials challenged preconceptions or reinforced them - and in their analysis found correlations between personality traits and political beliefs and the effect of the editorial. Durmus and Cardie [2018a] analyzed online debates[4] and concluded that self-provided demographic information and political and religious beliefs should be controlled for in persuasiveness prediction. Durmus and Cardie [2019b,a] show that debate experience and similarity between debater and audience have significant effects, although online debate skills can improve over time [Luu et al., 2019]. Finally, Longpre et al. [2019] find that *a priori* undecided users are more likely to be influenced by demographic factors and prior beliefs than decided users that change their stance on a topic. Unlike these approaches, we assume that posters participating in Change My View are open to changing their beliefs and any personal information is captured in their post history only.

While one approach to persuasion is to model self-reported belief change, another approach is to use a measurable definition of change based on the action the audience takes in response. Yang et al. [2019] study crowdfunding platforms,[5] labeling persuasion *strategies* such as concreteness or emotion and taking a semi-supervised approach to predict strategies and persuasiveness jointly, and find that 1) joint modeling is key and 2) concreteness and emotion are effective strategies, among others. Wang et al. [2019] conduct an experiment on Mechanical Turk where workers take a personality test, participate in a dialogue, and attempt to convince each other to donate to charity. The authors label persuasion strategies such as appeals to credibility or emotion and predict strategies using a neural model with additional features. They find that providing information about the charity and asking questions (but only when the recipient has the "openness" personality type) are effective techniques. Yang et al. [2015] predict users' brand preferences using personality traits and personal values and Ding and Pan [2016] find that framing, or emphasizing aspects of products

[4] `http://debate.org`
[5] `https://www.kiva.org/`

given personal information, is more effective for content selection in advertisements. Gao et al. [2019] study the factors in author responses in NLP conference proceedings and find that reviewers do change their scores, but mostly due to peer effects – trying to achieve consensus. While we leverage self-reported changes in belief, we observe some of the same helpful features.

A closely related area of research is the detection of situational influencers — participants in a discussion who have credibility in the group, persist in attempting to convince others, and introduce ideas that others pick up on or support [Rosenthal and Mckeown, 2017, Biran et al., 2012]. In particular, Rosenthal and Mckeown [2017] draw their approach from the idea of "weapons of influence" [Cialdini, 2005], which include reciprocation (sentiment and agreement components), commitment (claims and agreement), social proof (dialog patterns), liking (sentiment and credibility), authority (credibility), and scarcity (author traits). Morio and Fujita [2018a] predict influencers based on features derived from activity history and word specificity. Influencers also have an impact on the diffusion of information through social media. Tan et al. [2014] control for message content by leveraging the fact that Twitter users often retweet the same message with different wording and find that message length and informativeness are important factors for propagation. Das et al. [2016] find that messages with persuasive intent were more likely to be propagated if the emotional or logical components of a message were selected in a topic-specific way. Influence is also distinct from power [Prabhakaran et al., 2012a,b], which involves a hierarchy and can be predicted by stylistic features such as formality and entrainment [Danescu-Niculescu-Mizil et al., 2011b].

Conversation Quality Lastly, aspects of persuasive and high-quality arguments are shared with those of high-quality online conversations. Napoles et al. [2017] label comments and threads from Yahoo News articles for agreement/disagreement, constructiveness, attempts to persuade or argue, and tone (sarcastic/mean/sympathetic/funny/etc.), among other aspects, and how these properties interact overall to form good conversations. Srinivasan et al. [2019] find that content moderation can improve conversation quality, but Habernal et al. [2018b] note that even in a heavily-moderated forum such as Change My View, ad hominem attacks persist.[6] Similarly, Wulczyn et al. [2016] find that personal attacks in Wikipedia can not be eliminated simply by banning malicious or anonymous users. Furthermore, Cheng et al. [2017] observe that any user can become a troll, either due to other trolls or simply a negative mood. Recent work has thus attempted to forecast conversation derailment before it escalates [Zhang et al., 2018, Chang and Danescu-Niculescu-Mizil, 2019] so that moderators can be notified with sufficient warning. In our work, we assume that participants are arguing in good faith and that any derailment of the conversation would be captured in the overall lack of persuasiveness. Finally, other work has attempted to predict controversial discussions using sentiment [Wang and Cardie, 2014b] or network features [Hessel and Lee, 2019], whereas argumentative conversations are by nature controversial.

2.1.2 Form/Structure

The models of Toulmin [1958] and Freeman [1991] are the main inspirations for research in computational work on argumentation structure. The key components of the Toulmin model are the claim, backing, and (implicit) warrant. Freeman also considers the role of a proponent, who presents and defends their claims, and an opponent, who critically questions them [Lawrence and Reed, 2019]. Toulmin [1958] focuses on the role of argument components whereas Freeman [1991] focuses on the support and attack relations between them. Both approaches consider that an argument consists of a main claim, or conclusion, and can be inferred by premises, or reasons, for or against, which may also be supported or attacked by other premises [Wachsmuth et al., 2017a].

In practice, identifying the structure of an argument requires **segmentation** into argumentative units, identification of the role of the **propositions**, and identification of the **relations** between propositions. Often, these steps are conducted independently, rather than jointly, or assumptions are made (e.g. a sentence is an argumentative unit) to simplify the process.

Segments Identifying whether a segment, or an argumentative discourse unit (ADU), is argumentative or non-argumentative is often the first step in a pipeline before modeling argumentative

[6] although Budzynska and Reed [2012] claim that ad hominems are not always argumentative fallacies

components (claims and premises) and the relations between them. The boundaries of ADUs are often left up to the judgment of the annotator and may be a phrase, clause or series of clauses [Jo et al., 2019]. Stab and Gurevych [2017a] first model segmentation at the token level using a BIO tagging approach and then predict components and relations jointly using Integer Linear Programming. Ajjour et al. [2017] replicated their model in the domains of essays, news editorials, and web discourse and find that semantic features perform best in-domain but that structural features (e.g. position in sentence or paragraph) perform best across domains, although note that segmentation is overlooked as a reason for degradation of performance on the overall end-to-end task. Eger et al. [2017] elided the issue by modeling the entire process end-to-end and predicting ADUs and their role along with distance to the related ADU all at the token level. In a departure from common practice, Jo et al. [2019] claim that an intermediate step is required, the identification of *propositions*, statements that are either true or false, as ADUs do not fully capture anaphora or implicit claims in the form of questions. Finally, discourse connectives and other functional phrases may or may not be included in the annotated segment. Madnani et al. [2012] termed these functional phrases "shell" and showed that it could be predicted accurately using both supervised learning and a rule-based approach.

In our work, we label segments at the proposition level and evaluate our models using gold segmentation[Chakrabarty et al., 2019b,c], but consider propositions at the sentence level for extraction. We also leverage the role of the argumentative shell in fusing sub-arguments.

Propositions In the pipeline approach to argumentation mining, the next step in identifying the structure is to determine the role of the proposition (e.g. claim or premise). Early work considered this a separate step from relation prediction [Moens et al., 2007], although more recently researchers have pointed out that the argumentative role of a sentence depends on its context and may instead predict relations first [Carstens and Toni, 2015] or use a multi-task learning approach [Eger et al., 2017, Schulz et al., 2018].

However, even without context, models have been shown empirically to have reasonable success at identifying these functional roles as segments often contain discriminative phrases. Shnarch et al. [2017] develop GRASP - a pattern-learning approach for weak labeling of argumentative components. Nguyen and Litman [2015] use topic models to build better word lists for classifying claims and premises; their goal was to separate the argument "shell," from content. Rosenthal and McKeown [2012] consider classification as a binary task – whether a sentence contains an opinionated claim or not. Daxenberger et al. [2017] conducted a study of multiple datasets to identify similarities in the conceptualization of claims, and found that some lexical properties generalized across data sets. Similar to these approaches, we show that we can leverage linguistic patterns for distant labeling and transfer learning to predict argumentative components[Chakrabarty et al., 2019b,c].

Other work has incorporated the surrounding context. Carstens and Toni [2015] note that a segment may be factual without context but argumentative in context; they thus first predict the relation between segments, then the argumentative role. Nguyen and Litman [2016] incorporate context from a window around the segment in order to predict components (and later relations). Shnarch et al. [2018] model whether a sentence is a context-dependent claim or evidence for a given topic. Stab and Gurevych [2014a,b] annotate and then predict argument components as part of an end-to-end pipeline for persuasive essays. Habernal and Gurevych [2017] predict argumentative components at the token level in web discourse using BIO tagging. While we also use contextual information when fine-tuning a model using transfer learning, we predict components given a single proposition [Chakrabarty et al., 2019c].

Relations The final step is to predict an argumentative relation given a pair of segments, which is often distinct from component classification, which involves predicting a label for a single proposition or other segment. Biran and Rambow [2011b] predict whether a segment is a justification given a claim. Cocarascu and Toni [2017] predict a 3-way classification task of whether a segment supports, attacks, or has no relation with another segment. Other work has modeled argumentative relations that enforce a graph structure instead of assuming a tree structure. Niculae et al. [2017]

use factor graph models to enforce structure constraints. Gemechu and Reed [2019] build graphs of arguments by exploiting similarity between concepts and aspects expressed in text. Somasundaran et al. [2016] model relations by linking sentences with shared concepts and then use PageRank to extract relevant features for essay score prediction. Other work has examined how to represent relations between segments in dialogues. Aakhus et al. [2014] argues that disagreement arises in a discussion due to participants providing reasons for earlier problematic statements. Ghosh et al. [2014] subsequently label online discussions with the target of a post and the callout that responds to the target. In our work, we model both discourse and dialogue relations as a binary task - whether a relation exists between two segments [Chakrabarty et al., 2019c].

Lastly, as component and relation identification are closely linked, joint learning has recently gained prominence. Stab and Gurevych [2017a] model component types and argument relations using Integer Linear Programming. Potash et al. [2017] use pointer networks to model a sequence of components and relations. Peldszus and Stede [2015] apply minimum spanning tree decoding to enforce constraints on components and premises. Eger et al. [2017] predict components and relations at the token level using a neural tagging model. Lugini and Litman [2018] use neural multi-task learning for argumentative structure in classroom discussions. Morio and Fujita [2018b] jointly learn components and both intra-post and inter-post relations in online discussions using a constrained pointer network. While we model relations as part of a pipeline, we fine-tune a model on distant-labeled data and then use the same initial parameters for both components and relations [Chakrabarty et al., 2019c].

2.1.3 Content

Although identifying the role of a statement in an argument is inherently contextual, it may be desirable to identify the *intrinsic* properties of an argument, e.g. whether a statement is factual or emotional. This requires a deeper understanding of semantics rather than the approaches used for coarse-grained identification of components and relations. At the conceptual level, this distinction dates back to Aristotle's Rhetorics [Aristotle et al., 1954]. Aristotle considered that a good argument consists of the contextually appropriate combination of *pathos* (appeal to emotion), *ethos* (appeal to credibility), and *logos* (appeal to reason). More recently, Freeman [2000] created a taxonomy of proposition types, classified according to their verifiability (e.g. a factual statement compared to a personal one). Other work has further distinguished argumentative relations by classifying them according to their role, e.g. using argument *schemes* [Walton et al., 2008] such as causal reasoning, rather than simply predicting the presence of a relation between segments.

Intrinsic Properties of Arguments For modes of persuasion [Aristotle et al., 1954], the Internet Argument Corpus [Walker et al., 2012b] includes the distinction between fact and emotion based arguments. Oraby et al. [2015] distinguish "factual" and "feeling" arguments in this corpus using a pattern-based bootstrapping approach. Other researchers [Duthie et al., 2016, Duthie and Budzynska, 2018a,b] have developed a methodology for the retrieval of *ethos* in political debates and classification using neural networks. Higgins and Walker [2010] traced back *ethos*, *pathos* and *logos* as strategies of persuasion in social and environmental reports. Their definition of *logos* applies both to premises and claims, while we consider *logos* as referring to arguments only. Habernal and Gurevych [2017] have also included *logos* and *pathos*, but not *ethos*, among the labels for an argumentatively annotated corpus. Alliheedi et al. [2019] annotated rhetorical moves in scientific publications, including appeal to authority and background information, along with a new typology for semantic frames based on FrameNet [Ruppenhofer et al., 2006].

Other work has examined the acceptability of propositions [Freeman, 2000]. In recent work [Park and Cardie, 2014, Park et al., 2015, Park and Cardie, 2018], researchers annotated and predicted subjective/objective arguments in a rulemaking forum, distinguishing between unverifiable, verifiable non-experiential, and verifiable experiential propositions; this distinction allows for evaluating whether the supporting evidence is appropriate – reason, evidence, and optional evidence, respectively. Guggilla et al. [2016] use a deep learning approach with different embedding features based on words, dependencies, and factual indicators to predict types of claims on this same corpus and the Internet Argument Corpus [Walker et al., 2012b]. On a different dataset, Hua and

Wang [2017] label and predict types of supporting evidence – study, factual, opinion, and reasoning. Identifying factual claims is also a key step for fact-checking. Hassan et al. [2017] determine whether a claim is "check-worthy," i.e. whether it is nonfactual, unimportant factual, or check-worthy factual. Similar work is conducted in political debates [Patwari et al., 2017, Jaradat et al., 2018] and parliamentary proceedings [Naderi and Hirst, 2018]. In our work, we predict whether a statement is factual, but we assume that we are already given a verifiable proposition [Hidey and Diab, 2018].

The classification of propositions as verifiable or logical is also key for argumentation strategy. Al-Khatib et al. [2016b] annotate a news editorial corpus for decisions such as when to support an argument with statistics or when to tell an anecdote. In an analysis of the flow of strategy types, Al-Khatib et al. [2017] find that patterns are topic-dependent, for example that style editorials are anecdote-driven and environmental editorials begin and end with an anecdote but include statistics in the middle. Song et al. [2016] focus on the task of not just identifying anecdotes, but also recommending informative and interpretable ones. Although our work does not consider anecdotes, anecdotes include appeals to emotion and credibility, two of the modes of persuasion we study [Hidey et al., 2017]. On the other hand, we do develop new methods for fact-checking [Hidey and Diab, 2018].

As dialogues differ considerably from monologues, some work has considered linking speech acts, such as "challenging" or "promising," to argumentative dialogue using illocutionary structures drawn from *Inference Anchoring Theory* [Budzynska and Reed, 2011, Budzynska et al., 2014a,b, 2016]. Zhang et al. [2017] studied the rhetorical role of questions in parliamentary proceedings, finding that agreement is one function. Other work has looked at predicting where an opponent or audience might disagree, also known as "counter-considerations," to preempt possible objections by the readers [Peldszus and Stede, 2015]. Al-Khatib et al. [2018] also develop a model of Wikipedia discussions as progress towards determining the best speech act to take. In our work, we have decided to treat the speech acts of agreements/disagreements in dialogue as distinct types of claims since, depending on the semantics of the embedded proposition, they can express sharedness (or not) of interpretations as well as evaluations [Hidey et al., 2017].

Complex Argument Relations As with classifying roles and relations in identifying argument structure, while the intrinsic aspects of a claim or premise can be inferred in isolation, considering the context is often key in interpretation and prediction of arguments. This results in a distinction between the argument as a product, where every claim and premise is explicit, and the argument as a process, where implicit context is needed. By using argumentation schemes [Walton, 1995], the full argument, including implicit reasoning, can be reconstructed, e.g. if an appeal to expert opinion is used, the implicit reasoning is that if the expert has a stance on a subject in their area of expertise then the subject is good/bad [Reed and Walton, 2003]. Reed and Walton [2007] and Walton et al. [2008] consider a taxonomy of ninety-six argument schemes, including analogies, appeal to authority, logical reasoning, and causation. Lawrence and Reed [2015] focus on predicting two of these types, expert opinion and positive consequences, and find that using surrounding context is key to success. Lawrence et al. [2016] further find that by classifying individual proposition types, it is possible to reconstruct the structure of the argumentation schemes. In a domain such as genetics research, where language is more complex than online debates, argumentation schemes have also been shown to be identifiable [Green, 2015]. Musi et al. [2018] annotate eight types of argument schemes using a different set of labels [Musi et al., 2016] on the Microtext corpus [Peldszus and Stede, 2016a] and demonstrate a mapping between discourse relations and their annotations. They find that causal reasoning is the most common type in their corpus; consequently we focus specifically on causal reasoning and develop models to identify causal relations [Hidey and McKeown, 2016].

2.1.4 Realization

While structure and content are necessary global and local properties of an argument, ultimately the realization of the underlying concepts in terms of style and framing is necessary. While other aspects of realization, such as sarcasm [Justo et al., 2014, Oraby et al., 2015, Khodak et al., 2018] occur in online debates, we assume conversations are in good faith and that where sarcasm is present it is used as a rhetorical device or for disagreement. Furthermore, we focus entirely on written text, while recent work has examined recorded speech as a medium for debate [Mirkin et al., 2018a,b].

Style Wang et al. [2017] find that style plays an important role in debate outcomes, given latent content. Danescu-Niculescu-Mizil et al. [2011a] verify in Twitter the long-held theory of accommodation, where participants in a discussion coordinate on word choice, syntax, and length, among other aspects. Other work has examined accommodation in disagreements, finding that participants are more similar when they agree than when they disagree [van der Pol et al., 2016], measuring style using functional markers. Sridhar and Getoor [2019] estimate the causal effects of tone using linguistic and sentiment markers, finding that factual and asserting tones have the most impact. Durmus and Cardie [2018b] find that linguistic expression varies significantly when accounting for gender and stance in online debates. In our work, we distinguish between content and style by using overlap between sets of words [Hidey and McKeown, 2018] and partition content and style for selection and re-writing of content.

Framing Chong and Druckman [2007] define framing as "when (often small) changes in the presentation of an issue or an event produce (sometimes large) changes of opinion." Researchers have thus analyzed and attempted to identify framing in arguments. Choi et al. [2012] find that in GMO debates, hedges occur more frequently in popular text than in scientific articles. Cano-Basave and He [2016] leverage semantic frames for distant labeling of arguments in political debates. Musi and Aakhus [2019] study the use of semantic frames in fracking debates, finding that evidence and reasoning are among the most common. Ajjour et al. [2019] and Hartmann et al. [2019] use an argument-specific set of frames, such as economics, feasibility, and safety, with Ajjour et al. [2019] predicting frames using multi-task learning and Hartmann et al. [2019] clustering frames in a latent semantic space. August et al. [2018] also use a targeted set of frames and find that alternative slogans in advertisements improve self-selection for experimental online studies. While we use semantic frames for predicting various argumentative tasks [Hidey and McKeown, 2018], we do not study the effects of different choices of frames.

2.1.5 Annotation

Aspects of argumentation have been annotated for many different types of discourse, including persuasive essays [Stab and Gurevych, 2014b], online comments [Biran and Rambow, 2011a] and other web discourse [Habernal and Gurevych, 2015], Wikipedia talk pages [Biran and Rambow, 2011b], micro texts [Peldszus and Stede, 2013b], and various other genres [Moens et al., 2007].

The manual extraction and identification of arguments and components is a time-consuming process often requiring skilled annotators, in one case 7,000 hours of work for a single debate [Lawrence and Reed, 2019]. Furthermore, domain expertise is often required in the case of scientific [Green, 2014, 2018] or legal [Moens et al., 2007] arguments. Results are mixed as to the level of complexity that non-experts are capable of annotating. Walker et al. [2012b] found that in creating the Internet Argument Corpus, Mechanical Turk workers were able to annotate agreement/disagreement and stance with high inter-annotator agreement but with lower agreement for fact/emotion, attacking, or sarcastic arguments. Peldszus and Stede [2016a] found that trained annotators could annotate the structure of "microtexts" reliably with high inter-annotator agreement. However, Miller et al. [2019] find that by having annotators iteratively identifying the main claim, claims, and premises, minimal training is required. Lavee et al. [2019a] also find that crowd workers are capable of annotating long texts for claims and that providing the full context is key. In other work, Habernal et al. [2017] find that *gamification* is an effective approach to labeling argumentative fallacies.

Regardless of whether experts or non-experts are used, annotation is time consuming and thus researchers have turned to distant supervision approaches. The advent of the Internet has provided new forums for disagreement and the data contained therein thus provides the opportunity for modeling new tasks computationally. Habernal and Gurevych [2015] perform semi-supervised argument mining by combining lexical and other features with distance features derived by clustering unlabeled debates. Al-Khatib et al. [2016a] predict whether text is argumentative by distant labeling sentences from an online debate portal[7] using the site metadata. Tan et al. [2016] first noticed that the Change My View subreddit provided self-labeled data for persuasion. Finally, domain-specific idiosyncrasies can be taken advantage of, as in the case of biomedical research articles, where if the title contains a tensed verb it is often the main claim of the argument [Graves et al., 2014]. We use both expert annotators [Hidey et al., 2017] and crowdsourcing [Hidey and McKeown, 2016, Hidey et al., 2017], but we also make extensive use of distant supervision [Hidey and McKeown, 2018, Chakrabarty et al., 2019b,c] to leverage data-hungry neural models.

2.2 Generation

While the analysis of arguments consists of many different subtasks, the generation of arguments has additional challenges. The task of generating arguments is associated with the field of text-to-text generation, where the input may range from a sentence to a long document and the goal is to produce text as an output. Some examples include machine translation [Hutchins and Somers, 1992], paraphrase generation [Bannard and Callison-Burch, 2005], or summarization [McKeown, 1982]. In these cases, the input and output are highly correlated and thus *parallel* corpora are beneficial. In the case of argument generation, the input may be a topic, claim, or an entire argument, and the output may be an argument or one of its components (e.g. a claim). In this case, the task is more similar to creative tasks such as dialogue [Ritter et al., 2011] or narrative [Callaway and Lester, 2001] generation, where the goal is, in general, to generate an appropriate response rather than a semantically equivalent one. On these tasks, there are thus different challenges due to the difference in goals. First, while neural models have made significant strides on tasks such as machine translation using maximum likelihood estimation [Sutskever et al., 2014], other work has shown that this approach leads to the generation of generic responses such as *"I don't know"* [Li et al., 2016a] for tasks such as dialogue generation. Second, Reed et al. [2018] showed that in hierarchical tasks that depend on sentence planning and discourse structure, end-to-end models

[7] `http://idebate.org`

struggle to reproduce these underlying operations without explicit supervision. Finally, other work has shown that these models have a limited understanding of semantics, either overfitting to divergences for machine translation [Carpuat et al., 2017] or hallucinating incorrect information not present in the input text [Kryscinski et al., 2019] for summarization; constraining models to have a specific type of output is thus more difficult.

To address these challenges, researchers have found that better data can reduce semantic divergences in machine translation [Vyas et al., 2018] or have turned to approaches such as using different objective functions to promote diversity in dialogue [Li et al., 2016a,b]. Recently, researchers have shown that pre-training large neural models on large datasets can produce long coherent and fluent outputs with diverse responses (e.g. GPT-2 [Radford et al., 2019]), but Kassner and Schütze [2019] found that pre-trained models still have difficulty with factual knowledge, in particular distinguishing between statements such as "Birds can fly" and "Birds cannot fly." To improve hierarchical generation, Kiddon et al. [2016] use a "neural checklist" to improve global coherence of recipe generation, whereas Zhai et al. [2019] find that this approach is very data-heavy and a hybrid approach generating a symbolic structure representation before a neural realizer improves narrative generation. Similarly, See et al. [2019] find that large pre-trained models such as GPT-2 [Radford et al., 2019] generate more repetitive and under-diverse narratives than a state-of-the-art hierarchical model for story generation [Fan et al., 2018]. On other tasks that require specific constraints, researchers have found that rule-based approaches sometimes address constraints better than neural models, e.g. for pun generation [He et al., 2019].

The desirable properties of an argument align with these three challenges: 1) novel content that is not correlated with the input [Tan et al., 2016, Wachsmuth et al., 2018b] 2) global structure, i.e. a conclusion supported by premises [Wachsmuth et al., 2017a] and 3) semantically coherent, task-oriented output, i.e. the argument should, at minimum, have the desired stance, and, ideally, have appropriate properties such as emotion. Many approaches to argument generation have thus investigated several of the key subtasks listed by Gatt and Krahmer [2017] as needed for generation: content determination (or selection), text structuring (or planning), and lexicalization (or realization). **Content determination** may be done in a retrieval-based or generation-based manner. Retrieval-based methods involve re-using past arguments or finding supporting evidence for a claim. Generation-based methods may generate novel concepts abstractively or realize an argument given retrieved evidence. In an argument creation system, there may also be a **planning** step that selects and orders arguments. Lastly, the **realization** of the argument may be constrained to have desirable semantic properties such as emotion or the correct stance and may involve re-writing of retrieved arguments or be part of an end-to-end system that performs all three steps jointly. We thus discuss work in content selection in Section 2.2.1, argument planning in Section 2.2.2, and argument realization in Section 2.2.3.

2.2.1 Content Selection

Aharoni et al. [2014], Levy et al. [2014], and Rinott et al. [2015] introduced a dataset and models for the detection of context-dependent claims and evidence relevant to a given controversial topic given a small set of Wikipedia pages, retrieved using the approach of Roitman et al. [2016]. Similarly, Stab et al. [2018b] and Stab et al. [2018a] retrieve documents from the Web and predict whether a sentence in relation to a topic contains a supporting or attacking argument, or is non-argumentative; Trautmann et al. [2019] later take a fine-grained approach by finding propositions within sentences instead. Wachsmuth et al. [2017d] and Wachsmuth et al. [2017e], on the other hand, retrieve entire arguments for and against a particular stance. They continue this work by retrieving counter-arguments from an online debate forum by using features for word and topic similarity [Wachsmuth et al., 2018b] and finding that the best counter-arguments are simultaneously novel and similar to the original argument. In other work, researchers retrieved previous arguments using Siamese networks [Le et al., 2018] or topic-specific clustering [Rakshit et al., 2017]. Other work has looked at retrieving sentences (or claims/premises) directly from a corpus, instead of a document or argument retrieval step. Levy et al. [2017] directly retrieve claims from Wikipedia, improving performance by collocating words in a "Wikification" step and later using distant-labeled data to re-rank sentences using neural networks [Levy et al., 2018] and applying their work to a large expanded document collection that also includes other sources such as newspaper articles [Ein-Dor et al., 2019], allowing for searching for matching claims directly in speeches [Lavee et al., 2019b]. More recently, there has been a focus on generalization - human debaters often resort to general arguments that can be applied regardless of topic and so Bilu et al. [2019] introduce an approach for arguing from “first principles,” a dataset for general-purpose rebuttals independent of topic [Orbach et al., 2019], and an approach to expand the boundaries of discussion by relating Wikipedia concepts according to lexical-semantic relations such as hypernyms [Bar-Haim et al., 2019]. While these approaches rely on predicting relations between propositions or identifying entire arguments, we take a hybrid approach and instead only predict the relation between the main claims of two arguments and then extract the most persuasive sub-components.

2.2.2 Planning

Generating an argument may also include an separate planning step. Zukerman et al. [2000] use a formal model of argumentation to propose valid arguments. Other work has also used formal models for rule-based argument planning [Guerini et al., 2003, 2004, 2007], but also with a model of the cognitive state of the recipient in persuasion. Aouladomar and Saint-Dizier [2005] use a template-based approach to generate procedural texts given data labeled for argumentative discourse acts. Green [2006] developed a discourse planner for the generation of biomedical arguments. Other work focused on the ordering of sentences, first identifying claims and premises and then linking them using causal relations [Reisert et al., 2015, Yanase et al., 2015, Sato et al., 2015]. Wachsmuth et al. [2018a] first demonstrate that when experts follow the same rhetorical strategy they have higher agreement on the selection and ordering of arguments; El Baff et al. [2019] then build on this work with a computational approach to argument synthesis, approaching it as a language modeling task where argumentative discourse units are the core segments. Hua et al. [2019a] incorporate a planning component to distinguish between function and content sentences in an end-to-end model; they later incorporate latent style into their sentence-level planner [Hua and Wang, 2019]. In contrast with these approaches, we conduct planning at the *paragraph-*

level by ordering for global coherence– we assume that paragraphs are locally coherent and already contain relations between propositions.

2.2.3 Realization

Finally, an argument may be realized as part of an end-to-end system or re-written given output from a previous step. Park et al. [2019] generate claims in response to a given claim, using a diversity penalty to promote different perspectives. Atkinson et al. [2019] generate *explanations* for why an argument is persuasive, finding that additional word embedding features improve performance. Wang and Ling [2016] trained neural abstractive models to summarize opinions and arguments. Recent work involved generating argumentative dialogue by maximizing mutual information [Le et al., 2018]. Additional work involved generating opinions given a rating of a product [Wang and Zhang, 2017b]. Bilu and Slonim [2016] and Bilu et al. [2015] use a template-based classifier to combine retrieved topics and predicates and negate retrieved claims, respectively. Other researchers generated political counter-arguments given retrieved external evidence [Hua and Wang, 2018]. Hua et al. [2019b] later improve their approach by incorporating a planning step and then introducing latent variables representing style [Hua and Wang, 2019]. Lastly, a related task is that of extractive summarization, where Wang et al. [2014] select the most representative opinionated sentences using sub-modular optimization and Misra et al. [2017] train supervised neural models to identify important arguments in online dialogues. Our work assumes arguments are mostly fully realized, and instead we perform edits to the argumentative shell [Madnani et al., 2012] to improve coherence.

Other research similar to our work includes controlled text generation, which may include changing semantic aspects such as stance or sentiment. Hu et al. [2017] jointly train a variational auto-encoder and a discriminator to generate sentences with a specific polarity; later work improved on this approach by also conditioning on an embedding for the target polarity [Lai et al., 2019] and by training on noisy data from corrupted similar sentence pairs [Kruengkrai, 2019]. Other work used an adversarial training objective to transfer polarity between a set of unaligned sentences [Shen et al., 2017]. In recent work [Prabhumoye et al., 2018], researchers used back-translation and adversarial techniques to transfer sentiment and political stance. Li et al. [2018] find that given an attribute (e.g. sentiment), it is possible to retrieve similar sentences, delete attribute phrases, and generate new phrases conditioned on an attribute; later work leveraged large transformer models on this same task [Sudhakar et al., 2019]. In work similar to ours, Chen et al. [2018b] learn to flip the bias of news headlines according to political beliefs using neural models. We take a similar approach, but instead leverage distant-labeled social media data to generate a new claim with contrastive meaning to a given claim [Hidey and McKeown, 2019].

2.3 Semantic Relations

Understanding semantic relations between argumentative components is necessary for a deeper understanding of argumentation. This understanding is key in subareas within argumentation such as the detection of implicit warrants [Habernal et al., 2018a], stance [Sobhani et al., 2015], agreement/disagreement [Abbott et al., 2011], and argument schemes [Walton, 1995]. On coarse-grained argumentative relation tasks, the goal is often to predict if there is a contrastive relation between a pair of segments (often framed as a three-way task, where the pairs may also be unrelated). This occurs in the detection of agreement/disagreement and stance (which may be pro, con, or neutral). We thus discuss work on contrastive meaning in Section 2.3.1. On fine-grained tasks such as the prediction of argument schemes, understanding discourse relations such as causality is important. In a dataset of micro-texts [Peldszus and Stede, 2016a], researchers found that the "cause" discourse relation was correlated with the "causal" argument scheme [Musi et al., 2018]. Other work has found that understanding discourse relations is key to understanding argumentative structure [Cabrio et al., 2013]. We thus discuss work in causal relations and discourse in Section 2.3.2

2.3.1 Contrastive Meaning

Understanding contrast and contradiction is key to argumentation as it requires an understanding of differing points-of-view. Previous work investigated the detection of different points-of-view in opinionated text [Al Khatib et al., 2012, Paul et al., 2010]. Abbott et al. [2011] predict agreement or disagreement between posts in online debates, finding that sarcasm and markers such as *"I know"* were correlated with agreement whereas markers such as *"actually"* were correlated with disagreement. Misra and Walker [2013] build on this work in a topic-independent way, theorizing that cue words and phrases indicate the speech act of disagreement, rather than the relation between the posts. Wang and Cardie [2014a] build a lexicon for agreement and disagreement by assuming that words have a certain polarity within a topic and are therefore indicative of sentiment. Rosenthal and McKeown [2015] leverage thread structure as distant labels and as features for predicting agreement, disagreement, or neither. Unlike these approaches, Chen et al. [2018a] model the interaction between posts and Sridhar et al. [2015] create a joint model of disagreement and stance for global consistency.

Contrast also appears in the study of stance, where the opinion towards a target may vary. The SemEval 2016 Stance Detection for Twitter task [Mohammad et al., 2016] involved predicting if a tweet favors a target entity. The Interpretable Semantic Similarity task [Agirre et al., 2016] called to identify semantic relation types (including opposition) between headlines or captions. Target-specific stance prediction in debates is also addressed [Anand et al., 2011], with Walker et al. [2012a] and Sridhar et al. [2014] leveraging dialogic properties to predict whether a view has a "pro" or "con" stance. Sobhani et al. [2015] demonstrate a link between argument mining and stance classification, where the authors cluster arguments and use the derived features for argument prediction and stance classification. Bar-Haim et al. [2017a] create a dataset of claims from debates and retrieved claims from Wikipedia, with binary labels for stance. They treat the task as one of targeted sentiment analysis, extracting the target and predicting its relative polarity by using a lexicon and accounting for contrast; they build on this work by building a targeted lexicon [Bar-Haim et al., 2017b]. Boltužić and Šnajder [2017] represent relations between claims as logical relations between concepts and find that their structured approach improves over a text-based baseline on stance detection. Durmus et al. [2019b] similarly represent relations between claims as a tree with the most general claim at the root, but combine text-based representations with this tree structure for improved results. Other work has shown that in certain domains, predicting stance towards a topic, such as vaccination, is especially difficult [Skeppstedt et al., 2017, 2018]. On the application side, Toledo-Ronen et al. [2016] leverage the stances of experts on topics, as appeal to authority is a common argumentative strategy and accounting for this may help with stance detection. Although stance is not a primary focus of our work, we leverage prior work in stance detection to retrieve prior arguments. Furthermore, stance is essential for persuasive counter-argument generation, as in order to change someone's mind the counter-arguments presented should not take the same view as the argument they are responding to.

While contrast and contradiction may occur due to the difference in subjective opinions, contradictions also occur due to the difference in perceived and objective reality. Fact-checking can be viewed as stance toward an event, resulting in research on politician statements [Vlachos and Riedel, 2014], news articles [Pomerleau and Rao, 2017], and community forums [Mihaylova et al., 2018, 2019]. The largest such dataset is the FEVER dataset [Thorne et al., 2018], resulting in around 150,000 claims and a shared task to retrieve evidence from Wikipedia and predict whether the claim is supported, refuted, or there is not enough information to determine. While the FEVER dataset used claims created by annotators and was limited in domain, Hanselowski et al. [2019] mine claims and evidence from the Snopes fact-checking site. Similarly, Augenstein et al. [2019] mine 26 fact-checking websites for claims and evidence and find that joint learning across datasets improves performance. In this work, we present experiments on the FEVER dataset, introducing a new, related dataset as well. Detection of deception encompasses both detection of fake news and intentionally misleading reviews [Ott et al., 2011, Li et al., 2014]. Cocarascu and Toni [2016] find that incorporating argumentative structure improves performance on deceptive review detection and Kotonya and Toni [2019] find a similar approach helps for fake news.

Contradiction has also become a part of the standard natural language inference (NLI) paradigm.

Initial work in rich textual entailment involved predicting a binary task – entailment or not [Dagan et al., 2005]; and the relation between entailment and argumentative relation prediction was demonstrated in the work of Cabrio and Villata [2012]. Later work [de Marneffe et al., 2008, Ritter et al., 2008] examined the presence of contradictions, where a contradiction is defined as a pair of sentences that are extremely unlikely to be simultaneously true, with the former investigating different types of contradictions from negation or antonyms to phrasal or structural differences and the latter those that can be expressed with functional relations. Most NLI datasets now label contradiction, entailment, or neutral [Bowman et al., 2015, Williams et al., 2018, Poliak et al., 2018a]. The increase in these resources with contrast and contradiction has also resulted in new methods for word embeddings with contrastive meaning [Chen et al., 2015, Nguyen et al., 2016, Vulić, 2018] or compositional representations containing knowledge of contradiction [Conneau et al., 2017]. The utility of these datasets in relation to argumentation was demonstrated on the argument reasoning comprehension task [Habernal et al., 2018a] – predicting which of two implicit warrants correctly links a claim and premise. The winning system used models pre-trained on the Stanford Natural Language Inference dataset [Bowman et al., 2015], using the Enhanced Sequential Inference Model [Chen et al., 2017b] to obtain contextual representations for sentence pairs [Choi and Lee, 2018]. Recent developments have also led to large models pre-trained on a language modeling task and fine-tuned on these datasets, demonstrating significant improvement [Peters et al., 2018, Devlin et al., 2019] and leading to transfer learning approaches in argumentation as well. These pre-trained language models have already proven useful for argumentative tasks such as stance detection [Durmus et al., 2019a] and fact-checking [Stammbach and Neumann, 2019], among others. In our work, we leverage pre-trained models that have demonstrated excellent performance on NLI tasks [Howard and Ruder, 2018, Devlin et al., 2019] for argument component and relation prediction [Chakrabarty et al., 2019b,c], fact-checking, and stance detection.

However, although these models have achieved great success on these datasets, some recent research has called into question the ability of these models to truly learn semantic relations; Poliak et al. [2018b] show that a "hypothesis-only baseline," where only one side of the pair is considered, achieves between a 5% and 35% absolute improvement over a random baseline on several datasets. Similarly, Niven and Kao [2019] show that on the aforementioned argument reasoning comprehension task, BERT learns to predict when a warrant contains the word "not". When the dataset is balanced by including an equal number of examples with this word, BERT achieves no better than the random baseline. This line of work is part of a larger framework of creating and fixing "adversarial" examples [Jia and Liang, 2017, Wallace et al., 2019, Ribeiro et al., 2018, Glockner et al., 2018, Nie et al., 2019, Alzantot et al., 2018, Ren et al., 2019], suggesting that these powerful models are often just picking up on lexical cues and lack the deep understanding of semantics needed for fine-grained argumentation tasks. We also create adversarial datasets for fact-checking and provide a system that makes progress towards addressing these deficiencies.

2.3.2 Causal Relations

While the performance of models on tasks related to contrast/contradiction may be overstated, detecting causal relations is regarded as a difficult task [Roze et al., 2019]. Causal relations are represented in both rhetorical structure theory (RST) [Mann and Thompson, 1988], which defines a global structure, and Penn Discourse Tree Bank (PDTB) theory [Prasad et al., 2008], which defines a shallow discourse semantics between segments. To some extent, their analyses correspond with each other, as researchers have demonstrated that it is possible to map between these formalisms with high accuracy [Scheffler and Stede, 2016, Bourgonje and Zolotarenko, 2019].

In argumentation, researchers have correlated argumentative structure and discourse relations, both for RST [Green, 2010, Peldszus and Stede, 2013a, 2016b, Stede, 2016, Accuosto and Saggion, 2019] and PDTB [Hewett et al., 2019] relations. Other work has leveraged explicit discourse markers [Stab et al., 2014, Eckle-Kohler et al., 2015] for labeling and predicting claims and premises; Madnani et al. [2012] noted that argumentative "shell" phrases often have a discourse function. Forbes-Riley et al. [2016] annotate PDTB relations in student persuasive essays and Zhang et al. [2016a] leverage predicted PDTB relations for the task of argumentative revision classification. At the semantic level, researchers have shown that PDTB relations are helpful for predicting argumentation schemes [Cabrio et al., 2013]. In work specific to causality, Musi et al. [2018] show that the "cause" discourse relation was correlated with the "causal" argument scheme. Although causal relations are under-explored for the task of argumentation, Song et al. [2014] showed that argumentation schemes (reasoning patterns) such as causal reasoning are positively correlated with essay quality, whereas generic schemes were negatively correlated. Other work has leveraged causal relations using excitatory or inhibitory templates such as *promoting* or *suppressing* to mine relations between claims and premises [Reisert et al., 2015, Yanase et al., 2015, Sato et al., 2015] or by mapping claims to "micro-structures" [Boltužić and Šnajder, 2017], i.e. logical forms such as *promotes(A,B)*, which capture causality.

Similar to contrast and contradiction, detection of discourse relations can be improved when lexical markers are present. Early work with the PDTB [Pitler and Nenkova, 2009] showed that discourse classes with explicit discourse connectives can be identified with high accuracy using a combination of the connective and syntactic features. Performance on *implicit* discourse relations, however, has lagged behind, in spite of attempts to leverage explicit markers to learn representations for discourse relations [Marcu and Echihabi, 2002, Blair-Goldensohn et al., 2007, Pitler et al., 2009, Biran and McKeown, 2013, Braud and Denis, 2016], to create augmented datasets [Rutherford and Xue, 2015, Lan et al., 2017], to transfer knowledge using a domain adaptation approach [Ji et al., 2015], or to jointly learn representations [Lan et al., 2017, Qin et al., 2017, Varia et al., 2019].

Similar approaches have been used for causal relations, as the presence of lexical markers such as implicit causality verbs [Rohde and Horton, 2010] or discourse connectives can aid in their detection. Sharp et al. [2016] create causal embeddings using explicit causal connectives for why-question answering and Son et al. [2018] adapt the approach of Pitler and Nenkova [2009] to predict explicit causal relations in social media and predict the span of the causal explanation as well. Using templates based on discourse connectives like "and/thus/but" combined with what they termed inhibitory and excitatory predicates, researchers extracted causal relations between events [Hashimoto et al., 2012] and later applied this process to why-question answering [Oh et al., 2013] and argument relation prediction [Reisert et al., 2015, Yanase et al., 2015, Sato et al., 2015].

Acknowledging the difficulties of implicit causal relation detection, we also leverage known explicit markers to learn new ones. In addition to the closed class of explicit markers, PDTB researchers recognize the existence of an open class of markers, which they call *AltLex* [Prasad et al., 2010]. Along these lines, researchers used the EuroParl parallel corpus to find discourse connectives in French using known English connectives and filtering connectives using patterns [Laali and Kosseim, 2014]. Additional work focused on specific causal constructions – events paired by verb/verb and verb/noun [Riaz and Girju, 2013, 2014] or using construction grammar to identify linguistic patterns [Dunietz et al., 2017]. Other researchers link frame semantics with Granger causality for time series extracted from news articles [Kang et al., 2017], as these approaches are verb-oriented and event-based. In our work, we develop methods for detecting causal relations using explicit markers and AltLexes [Hidey and McKeown, 2016]. We also mine claims from social media using our approach to provide suggestions for electricity saving.

Chapter 3: Data

We consider two primary sources of information for our experiments, leveraging social media and an online encyclopedia. First, social media is a key venue for opinionated discussion as the majority of Americans obtain information about current events from social media platforms [Matsa and Shearer, 2018]. We thus use the website Reddit.com as the nature of the website allows us to efficiently obtain dialogues with opinionated posts and metadata about those posts. Second, verifying information against a trusted, authoritative source is a key component to establishing credibility towards the goal of persuasion. For this reason, we use Wikipedia.com as it is one of the most trusted sources of information on the Internet [Jordan, 2014] and the information on Wikipedia is largely considered to be as accurate as traditional encyclopediae [Giles, 2005].

3.1 Reddit

In order to study argumentative discourse and dialogue in social media, we need corpora that reflect the phenomena referenced in Section 2 such as persuasion and stance. Due to the community nature of Reddit, discussions may be centered around certain topics or devoted specifically to debate, which provides a diverse set of opinions for analysis. Furthermore, Reddit has no character limit for posts or limit to the number of responses in a thread, thus providing both discourse and dialogue structure. Finally, Reddit metadata and acronyms provide many options for distant labeling. All these features make it possible to study aspects of argumentation.

Reddit as a source of data also provides a few advantages compared to other social media platforms: 1) the data is publicly available and provides a number of APIs for easily obtaining data (e.g. pushshift)[1] 2) retrieving the structure of a discussion thread is trivial, unlike other social

[1] `http://pushshift.io`

media sites such as Twitter and 3) it is the 5th largest website in the US[2] with millions of English-language comments written every day.

Reddit is organized into a number of different "subreddits." These subreddits are smaller, organized communities within the Reddit infrastructure and may range from millions of regular visitors (e.g. default subreddits such as /r/television or /r/sports) to only a few dozen. When a Reddit user makes a post, they do so in a particular subreddit and depending on the engagement, popular posts may receive hundreds of thousands of upvotes.

3.1.1 Change My View

One such subreddit is known as "Change My View." Change My View (CMV) is a discussion forum where users post their opinions on a topic and other users post counter-arguments to that opinion. The initiator of the discussion will create a title for their post (which contains the major claim of the argument) and then describe the reasons for their belief. Other users respond by posting arguments attempting to change the view of the initiator of the discussion. If the view of the original poster (OP) is successfully changed, they will indicate this by posting a response with a "delta" character (Δ), providing naturally labeled data.

Consider the example in Table 3.1. In this discussion, the OP "A" states her belief that borders between nations are just a social construct. The user "B" responds with her own argument that even though borders are not a natural occurrence, it is human nature to require this kind of organization. The OP "A" then responds with a delta and acknowledges that she doesn't have a legitimate counter-argument. The overall structure of the argument is clear: the user begins by introducing evidence, making a concession as a matter of politeness, and finally concluding with a summarization and rhetorical questions.

Due to the nature of this subreddit, this forum provides many advantages for studying naturally-occurring persuasive argumentative discussions for the properties of persuasion discussed in Section 2.1.1. Unlike other argumentative forums, topics are not predefined to be in a particular

[2] https://en.wikipedia.org/wiki/Reddit

User	Post
Title	CMV: my view is that nations are just lines on a map and not real or useful
A	Nations are just lines on a map and don't exist in reality, here's my reasoning: 1) No one can decide where a nation begins or ends. Everyone's conception of "the South" when talking about America for example, will include different states and regions than the next person. In Europe, Turks claim that Cyprus is part of their nation, while Greece claims that island. Both claim Constantinople. Similarly, ...
B	https://en.wikipedia.org/wiki/Social_fact There is a word for what you are describing. While I'd concede your point is potentially valid, using your line of thinking makes living as a human being really difficult ... social facts make living in a human society possible in the first place. While they might be technically no true/real in a certain sense of the word, they provide structure in an otherwise structureless world. What's better? Have some orientationen, even though it's technically wrong. Or live without any kind of point of orientation, in a structureless world?
A	I'm going to give you a delta because you totally nailed it with the definition and your third paragraph raises points I can't answer: Δ

Table 3.1: Truncated Discussion Thread in Change My View

coarsely-defined category (e.g. abortion or gun control) as with some of the debate forums discussed in Section 2.1.5 but rather offer more fine-grained and nuanced opinions.

Furthermore, because of the community culture, ad hominems and bad faith arguments are limited [Habernal et al., 2018b] and discussions in this subreddit are likely to have the aspects of conversation quality discussed in Section 2.1.1. Users are encouraged to post in good faith and have an open mind, while soapboxing and proselytizing is discouraged.[3] Visitors to the forum are provided with guidelines on improving debate skills and increasing awareness of argumentative fallacies.[4] Additionally, participants are encouraged to grant the benefit of the doubt and assume the strongest argument[5] and "continue the conversation" rather than win a debate. Also, users are encouraged to provide deltas if any aspect of their view changed, even for partial movement towards the counter-argument.

Finally, the subreddit is heavily moderated, encouraging well-written arguments.[6] Original posts are removed unless they contain more than 500 characters consisting of coherent content (Rule A), consist of a view held by the original poster (Rule B), have a non-inflammatory statement

[3] https://www.reddit.com/r/changemyview/wiki/index
[4] https://www.reddit.com/r/changemyview/wiki/argumentation
[5] https://www.reddit.com/r/changemyview/wiki/guidelines
[6] https://www.reddit.com/r/changemyview/wiki/rules

as a title that sums up the viewpoint (Rule C), express a non-neutral stance (Rule D), and have the original poster make a high-effort response to counter-arguments within 3 hours (Rule E). Comments are removed if they do not address an issue with the original post, e.g. if they simply express agreement (Rule 1), if they are rude or hostile (Rules 2 and 3), or if they do not contribute to the conversation, e.g. with memes or low-effort comments (Rule 5).

Deltas can be provided by any user and received by any user except the OP in a particular discussion. When providing a delta, a user must also provide an explanation for why their view changed. Deltas are removed for a number of reasons, including sarcastic or joke deltas or those awarded on another user's behalf, i.e. if the user feels they should have been awarded one (Rule 4).

All in all, this means that we can reliably infer that the original post has a title with an opinionated claim and evidence to back up the claim and the original poster is engaged with the responding posters, all responses to the original post are counter-arguments with the opposite stance and with a high effort to address the original argument, and when a user provides a delta this is a reliable signal that their view was changed.[7] This structure provides us with a number of potential tasks to study, including predicting which of two counter-arguments is more persuasive (as with work described in Section 2.1.1) and generating counter-arguments (Section 2.2).

3.1.2 Distant Supervision

While Change my View provides a fairly structured source of opinionated discussions, Reddit as a whole contains argumentative dialogue as well that we aim to utilize. Although Reddit lacks the moderation and community of Change my View that provides relatively high quality argumentative dialogue, we can leverage heuristics to obtain automatically-labeled, noisy data that acts as a proxy for several argumentative phenomena of interest. As Reddit contains billions of comments, given a reasonably high-precision method for distant supervision even a phenomenon with a very low base rate can provide us with millions of training examples.

[7]Conversely, while this means that false positives are very low, false negatives are still a problem, but due to Rule 4 these are minimized as well

CMV: A rise in female representation in elected government isn't a good or bad thing. According to this new story, a record number of women are seeking office in this year's US midterm elections. While some observers hail this phenomenon as a step in the right direction, I don't think it's good thing one way or the other: a politician's sex has zero bearing on their ability to govern or craft effective legislation. As such...
>I don't think it's good thing one way or the other: a politician's sex has zero bearing on their ability to govern or craft effective legislation Nobody is saying that women are better politicians than men, and thus, more female representation is inherently better for our political system. Rather, the argument is that...

Table 3.2: Two posts from the CMV sub-reddit where a claim is targeted by the response

Previous work has examined the use of "sarcasm hashtags" [Davidov et al., 2010] to leverage self-labeled examples to predict sarcasm. Similarly, we find that users self-label several interesting aspects of argumentation.

First, we note that markdown language[8] provides a way to highlight specific subtext in another user's post by prefacing it with the > character (in Reddit markdown, this appears as indented text). We refer to this as the "quote" feature. In CMV, posters use this feature to respond to specific points in an argument and provide a counter-argument, providing a naturally-occurring instance of an argumentative relation such as agreement or disagreement (or at the minimum, semantic relatedness). In the example in Table 3.2, the response contains an exact quote of a claim in the original post.

Second, we leverage the use of internet acronyms[9] as a method of automatic data collection [Hidey and McKeown, 2019, Chakrabarty et al., 2019b,c]. Specifically, we mine comments containing the acronyms IMO (in my opinion) or IMHO (in my humble opinion) and FTFY (fixed that for you). As these acronyms are commonly used and can be removed from the comment without changing the meaning, we can still learn linguistic patterns of arguments from the remainder of the comment, as compared to an automatic labeling method such as GRASP [Shnarch et al., 2017]

[8] `https://en.wikipedia.org/wiki/Markdown`

[9] `http://https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/what-do-all-these-acronyms-mean`

that uses n-gram patterns.

IM(H)O is a commonly used acronym with the only purpose of identifying one's own comment as a personal opinion. As Reddit does not have a character limit, the opinion is usually elaborated upon in the following sentences. This allows us to assume that a relation exists between the sentence containing IMHO and the following one. While a premise does not always directly follow a claim, it does so frequently enough that this noisy data should be helpful [Chakrabarty et al., 2019b,c]. In the following example, the poster makes a claim and follows it with an explanation.

> **IMHO, Calorie-counting is a crock what you have to look at is how wholesome are the foods you are eating.** <u>Refined sugar is worse than just empty calories - I believe your body uses a lot of nutrients up just processing and digesting it.</u>

FTFY is a common acronym meaning "fixed that for you" and is used to respond to another comment by editing part of the "parent comment." This allows us to use comments containing this acronym as distant-labeled examples of contrast. Consider the following example,

> I've been saying this for years. If you want to decrease abortions where the fetus is not at health risk of defect, increase wages, increase job opportunities for single women, increase child care and education, etc. The christain Taliban has never been pro-life, they're just **pro-fetus**.
>
> >The christain Taliban has never been pro-life, they're just ~~pro-fetus~~ **anti-woman**.
>
> FTFY

The poster uses the quote feature to select a specific part of the parent comment to respond to and then edits their response so that it has different point of view (a stronger negative framing of opponents of abortion). This provides us with labeled post-response pairs for disagreement that are also labeled at the token level for contrast [Hidey and McKeown, 2019].

Most commonly, FTFY is used for three categories of responses: 1) expressing a contrastive opinion (e.g. the parent is ***Bernie Sanders*** *for president* and the FTFY is ***Hillary Clinton*** *should* *be president*) which may be sarcastic (e.g. ***Ted Cruz** for president* becomes ***Zodiac killer** for president*) 2) making a joke (e.g. *This Python library really **piques** my interest* vs. *This really ***py*ques** my interest*), and 3) correcting a typo (e.g. *This **peaks** my interest* vs. ***piques***).

3.2 Wikipedia

Although the Wikipedia organization does not claim to be a reliable source, numerous studies have indicated that it is more reliable than static sources such as the Encyclopedia Britannica.[10] studies have even shown that people find Wikipedia to be more reliable than the news itself [Jordan, 2014], even though Wikipedia requires citations from sources such as news articles. Unlike the news articles or resources that Wikipedia is based on, Wikipedia offers a unique combination of a widely-viewed website that is also perceived to have a high degree of reliability, making it a suitable authority for our experiments on fact-checking [Hidey and Diab, 2018]. Establishing trust is one of the key components discussed in persuasion and argument quality (Section 2.1.1) and reference/citation is often considered one of the semantic types of arguments (Section 2.1.3). Consequently, arguments often cite Wikipedia as a source, as in our example in Table 3.1.

Furthermore, Wikipedia is the 10th most popular website in the world[11] with over 6 million articles in English.[12] Moreover, in addition to English Wikipedia, Simple English Wikipedia has over 100,000 pages as well. Studies have shown that Simple Wikipedia is not just a paraphrase of the English language articles in simpler language [Yasseri et al., 2012]. We thus leverage the correspondence between article titles to find English-language paraphrases, which we use for modeling causal relations [Hidey and McKeown, 2016]. In fact, there is new information in Simple Wikipedia that is not in the corresponding English Wikipedia article and there are even articles in Simple Wikipedia that are not in English Wikipedia. Thus, future work could take advantage of Simple Wikipedia for better coverage of topics and viewpoints. Similarly, while our research is primarily on English-language arguments, Wikipedia contains articles in many different languages.

[10] `https://en.wikipedia.org/wiki/ReliabilityofWikipedia`
[11] `https://en.wikipedia.org/wiki/Listofmostpopularwebsites`
[12] `https://en.wikipedia.org/wiki/Wikipedia:SizeofWikipedia`

Future work can then leverage information only available in other languages.

3.3 Limitations

While Reddit is a large website and thus contains ample data for modeling, we must consider that valid opinions may not be adequately represented or that undesirable opinions may be over-represented. Surveys of Reddit have indicated that the site is mostly white non-Hispanic (70%), male (more than 67%), young (64% under 30), or American (58%) [Sattelberg, 2019]. Furthermore, only 4% of American adults have used Reddit and most users of the site are viewers rather than posters. This suggests that the most opinionated users have their views represented while other perspectives are lacking. It is thus important to be mindful of this fact when applying techniques derived from Reddit data; while we may be able to find certain viewpoints and topics with high precision we may also have low recall for others.

Wikipedia has similar problems with coverage of viewpoints and topics. While the information present in Wikipedia is mostly accurate [Giles, 2005], only between 8.5% and 16% of the editors are female. This has led to issues such as the ratio of articles about men to articles about women being 3 times as large as other encyclopediae.[13] Other issues appear in the language used - articles about women tend to contain gendered, bias language [Wagner et al., 2015]. As with Reddit, we should keep in mind that Wikipedia is a high-precision resource but may have low recall for certain under-represented topics.

[13] https://en.wikipedia.org/wiki/GenderbiasonWikipedia

Chapter 4: Analysis of Arguments

In order to make progress towards the generation of effective arguments, interpretation of arguments is key. As discussed in Chapter 2, effective argumentation is determined by structure [Ghosh et al., 2016], content [Tan et al., 2016, Wachsmuth et al., 2018b], and realization [Wang et al., 2017], among other factors such as discourse relations and speech acts. While many of these aspects have been studied individually in diverse corpora such as persuasive essays or online debates, understanding the interaction between these different components is necessary to provide a persuasive argument.

We thus unify these aspects by conducting experiments in a persuasive online discussion forum, which allows us to study argumentation at both the *micro level* [Somasundaran et al., 2007], the intra-post relations between propositions within a single comment or essay, and the *macro level* [Bentahar et al., 2010], the inter-post relations across turns of dialogue. Section 3.1.1 described the Reddit discussion forum Change My View, which provides a naturally-labeled dataset for predicting persuasion. Then, by annotating dialogues in this persuasive discussion forum for structure and semantics and building predictive models of argumentation and persuasion, we can make progress towards understanding the interaction between form and content.

We first annotate argumentative dialogues for their structure (i.e. claims and premises and the relations between them) and semantics (i.e. appeals to logic, emotion, and credibility). We then conduct an analysis of the resulting data for the persuasive impact of their interaction [Hidey et al., 2017, Chakrabarty et al., 2019c]. Next, we use this dataset to build models of argumentation structure, first predicting claims and premises and then predicting when a relation is present, leveraging discourse and dialogue context for micro-level and macro-level models, respectively [Chakrabarty et al., 2019b,c]. Finally, we present an alternative approach where we predict the persuasiveness of an argument using neural models of intra-post and inter-post interaction [Hidey and McKeown, 2018].

As claims are the central component of an argument [Govier, 2010], in order to respond and refute the claims of an opponent's arguments, we first need to identify them. Furthermore, as our generation approach uses a retrieve-and-edit framework, a model of the structure of argument allows us to select and extract argumentative content. Finally, these models further our understanding of how form and content interact, which allows us to automatically understand and select effective arguments.

Section 4.1 describes the annotation scheme and process along with an analysis of the data for persuasive impact. Section 4.2 describes our models for detecting claims and premises and the relations between them at the intra-post or inter-post level, obtained by fine-tuning a language model on the appropriate discourse or dialogue context. Section 4.3 describes our neural model of persuasion that accounts for argument sequencing and semantics as well as interaction between posts.

4.1 Annotation of Social Media Dialogues

Recent work in argumentation mining and persuasion has explored the role of discourse form in persuasive essays [Stab and Gurevych, 2014a, Peldszus and Stede, 2016a, Ghosh et al., 2016] and the features that make online arguments persuasive [Habernal and Gurevych, 2016a, Tan et al., 2016]. We build on and extend previous work by unifying these aspects in a single corpus of online arguments annotated for 1) claims and premises and their interaction at the micro and macro level 2) conceptual semantic features and 3) persuasiveness. The combination of these aspects enables the analysis and automation of persuasive argumentation.

On these grounds, we propose and validate a systematic procedure to identify conceptual aspects of persuasion, presenting a three-stage annotation process on a sample of 112 threads from *Change My View* (Section 3.1.1). *Change My View* constitutes a suitable environment for the study of persuasive argumentation as the provision of deltas results in a naturally-labeled dataset for persuasive arguments. In the first stage, expert annotators are asked to identify claims and premises among the propositions forming the post [Hidey et al., 2017]. In the second stage, using crowdsourcing (Amazon Mechanical Turk), claims and premises are annotated with their semantic types [Hidey et al., 2017]. We conduct our annotation process in multiple stages because experts have traditionally been used to label the structure of arguments [Peldszus and Stede, 2016a] but crowd workers have demonstrated their capability on similar semantic tasks (e.g. labeling factual and emotional arguments [Walker et al., 2012b]). For premises, the semantic types are based on the Aristotelian modes of persuasion: *logos*, *pathos* and *ethos*, or any combination of the three [Aristotle et al., 1954]. For claims, we have considered two proposition types among those in Freeman's taxonomy [Freeman, 2000] that can work as claims since their truth is assailable, namely *interpretations* and *evaluations (rational/emotional)*. We have furthermore distinguished propositions expressing *agreement* and *disagreement* because they present an anaphoric function inherent to the dialogic nature of the corpus. Finally, in the third stage, we use the same expert annotators to label relations between claims and premises [Chakrabarty et al., 2019c]. Similar to previous work in monologues (e.g. persuasive essays), we label *intra-turn* relations (to support or attack one's own claim) in an argument within a single post, also known as the *micro-level* model of argument [Somasundaran et al., 2007, Stab and Gurevych, 2014b, Swanson et al., 2015, Feng and Hirst, 2011, Habernal and Gurevych, 2017, Peldszus and Stede, 2015]. Additionally, we annotate *inter-turn relations* (argumentative relations to support or attack the other person's argument) across posts from two different users. This is known as the *macro-level* model of argumentation (or dialogical model), which focuses on the *process* of argument in a dialogue [Bentahar et al., 2010] and has received less attention.

Figure 4.1 shows a thread structure consisting of multiple posts with argumentative components (main claims, claims or premises) and both intra- and inter-turn relations where the original poster (OP) changed their view.[1] In this example, R disagrees with the main claim *Patriotism is the belief that being born on one side of a line makes you better* by re-stating the claim with a different view. They go on to provide two examples that support this contrasting claim while also making

[1]Note that premises are labeled at proposition level and not clause level.

CMV: [Patriotism is the belief that being born on one side of a line makes you better.] **OP**
[**0:CLAIM**]
[…]

[I would define patriotism quite simply as supporting one's country, but not *necessarily* disparaging others.][**10:CLAIM_DISAGREEMENT:0_DISAGREEMENT**] **R**
[...]
[Someone who attempts to extract gain for themselves at harm to their overall country is acting unpatriotically][**12:PREMISE_LOGOS_PATHOS:10_SUPPORT**]
[Someone who assists another country that is in worse shape instead of assisting their own can still be a patriot, but also recognize significant need in other nations and decide to assist them as well.][**13:PREMISE_LOGOS_PATHOS:10_SUPPORT**]
[...]

[This is true][**16:CLAIM_AGREEMENT:12_13_PARTIAL_AGREEMENT**] but, **OP**
[I think, supporting the common good is also more important than supporting your country.]
[**17:CLAIM_RATIONAL_EVALUATION:12_13_PARTIAL_AGREEMENT**]

[Yes,][**18:CLAIM_AGREEMENT:17_PARTIAL_ATTACK**] but **R**
[the two are often one and the same,][**19:CLAIM_INTERPRETATION:17_PARTIAL_ATTACK**]
[especially when you live in a country as large as the U.S., most acts which support the common good generally support your country.][**20:PREMISE_LOGOS:26_SUPPORT**]

Figure 4.1: Example from Change My View at the End of the Three-Stage Annotation Process. Each proposition is labeled with an ID, the proposition type and the relation node ID and relation type.

an emotional appeal. Both OP and R then concede the other's point. OP mostly agrees with the examples R provided but makes an evaluative claim that supporting the common good is better. R responds by mostly disagreeing with OP that there is a distinction and provides another example.

We describe how we select data for annotation in Section 4.1.1. In Section 4.1.2, we explain the annotation scheme and process. We show that experts can reliably annotate claims and premises and the relations between them. We also show that crowd-workers can reliably annotate the semantic types of premises but obtain lower agreement for the semantic types of claims, providing a qualitative analysis to understand when low agreement occurred (Section 4.1.3). In Section 4.1.4, we study the correlations between types of argumentative components (premises and claims), as well as their position in the post. We further show that there are several significant differences between persuasive and non-persuasive comments as to the types of claims and premises.

The annotated data and subsequent analysis allows us to progress towards automatic identification of argumentative structure, thus enabling forward movement towards automatically detecting persuasive arguments. The annotated dataset is available on GitHub to the research community.[2]

4.1.1 Data

We use data from the Change My View forum described in Section 3.1.1 and used in previous work [Tan et al., 2016]. We extract dialogs from the full dataset where only the original poster and one responder interacted. If the dialogue ends with the original poster providing a Δ, the thread is labeled as positive; if it ends prematurely without a Δ, it is labeled negative. In order to ensure that we have a data set representative of persuasive impact, we select discussion threads where there is at least one positive and one negative dialogue. In other words, each original post should have at least two responses, with at least one discussion resulting in the original poster changing their view and at least one where the original poster remains unpersuaded. We thus obtain 39 discussion threads to be annotated, consisting of 49 positive dialogues and 63 negative dialogues.[3] In total, the dataset contains 380 posts/turns of dialogue for 2756 sentences.

4.1.2 Annotation Scheme and Methods

Claims and Premises In the first stage of the annotation process, the goal is to label claims and premises at the proposition level. We recruited 8 students with a background either in Linguistics or in Natural Language Processing to be annotators. Students were asked to read the guidelines and were given an example with gold labels. During a one-hour long training session they were asked to annotate a pilot example and comparison between their preliminary annotations and the gold labels was discussed. Each student annotated from a minimum of 5 to a maximum of 22 threads depending on their availability.

The guidelines provide an intuitive definition of claims/premises paired with examples. While

[2] `https://github.com/chridey/change-my-view-modes`

[3] The first version of our corpus [Hidey et al., 2017] consisted of 39 positive and negative threads to have a balanced dataset for analysis, whereas the second version [Chakrabarty et al., 2019c] was further annotated to obtain additional examples for training a model.

the definitions are similar to those provided in previous annotation projects [Stab and Gurevych, 2014b], we took as annotation unit the proposition instead of the clause, given that premises are frequently propositions that conflate multiple clauses (see Figure 4.1). We further instructed the annotators that when two propositions are linked by a connective (*"but," "because,"* etc.), the connective is not part of the labeling: e.g. *"Everybody should eat at least one fruit a day because it is healthy."*

We informed the annotators that an argument is a justification provided by the speaker/writer in support of a claim that is not self-evident. Arguments are put forward by speakers/writers in order to persuade hearers/readers to agree about the truth of a certain claim. For claims and premises, we provided the following definitions and examples:

- **claim**: proposition that expresses the speaker's stance on a certain matter. They can express predictions (e.g. *"I think that the left wing will win the election"*), interpretations (*"John probably went home"*), evaluations (*"Your choice is a bad one"*) as well as agreement/disagreement with other people's claims (*"I agree"*)/(*"I think you are totally wrong."*) Complex sentences where speakers at first agree and then disagree with other speakers' opinion (*concessions*) constitute separate claims (*"I agree with you that the environmental consequences are bad, but I still think that freedom is more important."*).

- **premise**: proposition that expresses a justification provided by the speaker in support of a claim to persuade the audience of the validity of the claim. Like claims, they can express opinions but their function is not that of introducing a new stance, but that of supporting one expressed by another proposition (e.g. *"John probably went home. I don't see his coat anywhere;" "Look at the polls; I think that the right wing will win the election."*)

Both claims and premises can be expressed by rhetorical questions, questions that are not meant to require an answer — which is obvious — but to implicitly convey an assertive speech act. Their argumentative role, thus, has to be decided in context: in the argument *"We should fight for our privacy on the Web. Don't you love that Google knows your favorite brand of shoes?"* the rhetorical question functions as an argument in support of the recommendation to fight for privacy.

Completely untagged sections mostly contain greetings, farewells, or otherwise irrelevant text. Thus, occasionally entire paragraphs are left unmarked. Furthermore, we left the title unannotated, assuming that it works as the original poster's major claim, while we are interested in the comments that could persuade the original poster to change his view. When the original poster's text starts with an argument, it is by default assumed to be in support of the title.

Types of Claims and Premises The second stage aims to label the semantic type of claims and premises using crowdsourcing. We used Amazon Mechanical Turk (AMT) as our crowdsourcing platform. Using the previous annotations of claims/premises, Turkers were asked to identify the semantic type of premises and claims. In contrast with previous work, we propose a fine-grained, limited-context annotation of semantic types of premises and of claims. Existing semantic classifications focus either on premises or on claims (Section 2.1.2). Other studies have tackled types of premise/claim combinations specific to a restricted set of argument schemes [Atkinson and Bench-Capon, 2016, Lawrence and Reed, 2016] or proposed a classification scheme of claims that accounts for the subjectivity/objectivity of the premises [Park et al., 2015], but neither make reference to modes of persuasion appealed through the premises.

For claims, the Turkers were asked to choose among four different choices. The distinction between interpretations and evaluations recalls Freeman's classification of contingent statements [Freeman, 2000]. We have decided to treat agreements/disagreements as distinct types of claims since, depending on the semantics of the embedded proposition, they can express sharedness (or not) of interpretations as well as evaluations. The provided definitions are:

- **interpretation**: expresses predictions or explanations of states of affairs (*"I think he will win the election."* or *"He probably went home."*)
- **evaluation**: the claim expresses a more or less positive or negative judgment. Drawing from the distinction made in sentiment analysis and opinion mining, [Liu, 2012] evaluations are sub-classified as:

  - **evaluation-rational**: expresses an opinion based on rational reasoning, non-subjective evidence or credible sources (*"His political program is very solid."* or *"He is a very smart student."*)
  - **evaluation-emotional**: expresses an opinion based on emotional reasons and/or subjective beliefs (*"Going to the gym is an unpleasant activity."* or *"I do not like doing yoga."*)

- **agreement** or **disagreement**: expresses that the speaker shares/does not share to a certain degree the beliefs held by another speaker, i.e. *"I agree that going to the gym is boring"* or *"you are right"* or *"I do not think that he went home."* or *"You are not logically wrong"* or *"I do not like your ideas"* or *"It may be true."*

For each claim, we showed the workers the entire sentence containing the claim as well as the following sentence. The interface and an example for the claim labeling task is displayed in Figure 4.2. The "Opinion Instructions" consist of the previously provided definitions and examples.

For premises, the Turkers were provided with the following labels:

- **logos**: appeals to the use of reason, such as providing relevant examples and other kinds of factual evidence (*"Eating healthy makes you live longer. The oldest man in the US followed a strictly fat-free diet."* or *"He will probably win the election. He is the favorite according to the polls."*)

- **pathos**: aims at putting the audience in a certain frame of mind, appealing to emotions, or more generally touching upon topics in which the audience can somehow identify (*"Doctors should stop prescribing antibiotics at a large scale. The spread of antibiotics will be a threat for the next generation."* or *"You should put comfy furniture into your place. The feeling of being home is unforgettable."*)

- **ethos**: appeals to the credibility established by personal experience/expertise (*"I assure you the consequences of fracking are terrible. I have been living next to a pipeline since I was*

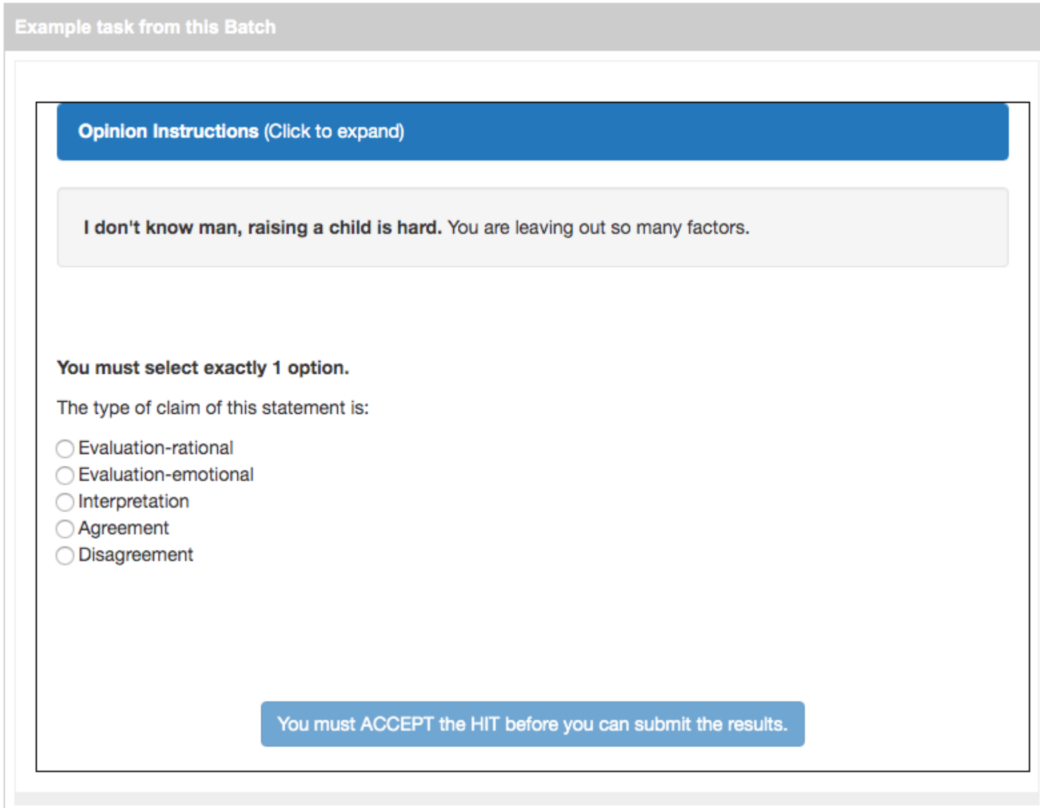


Figure 4.2: Amazon Mechanical Turk Claim Task

a child." or *"I assure you the consequences of fracking are terrible. I am a chemical engineer."*) as well as title/reputation (*"I trust his predictions about climate change. He is a Nobel Prize winner."* or *"I trust his predictions about climate change. They say he is a very sincere person."*)

For each premise, we showed the Turkers the entire sentence containing the premise, the following sentence, and all prior text up to and including the sentence containing the previous claim (see Figure 4.3, "Analysis Instructions" again includes the given definitions and examples). In operational terms, the workers were asked to select *true* for the persuasion mode used and *false* for the ones that were not applicable. They were given the choice to select from 1 to 3 modes for the same premise. If the workers did not select any modes, their HIT was rejected.

Each HIT consisted of 1 premise or 1 claim classification task and we had 5 Turkers for each task. The Turkers were paid 5 cents for each HIT in accordance with the U.S. federal minimum

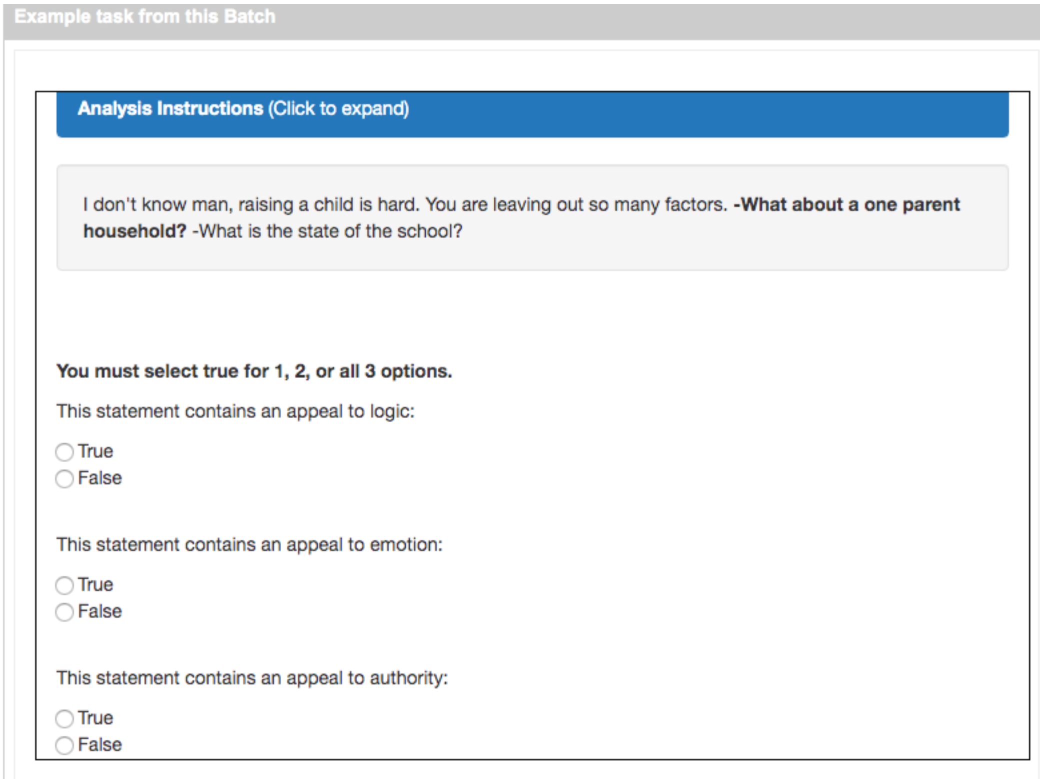


Figure 4.3: Amazon Mechanical Turk Premise Task

wage and we required the Turkers to have obtained the Masters level.

Argumentative Relations In the third stage, we use the same expert annotators to label the argumentative relations (both **inter-turn** and **intra-turn**) among the labeled propositions from stages 1 and 2 on the same posts they previously annotated.

As in prior work [Morio and Fujita, 2018b], we restrict **intra-turn relations** to be between a premise and another claim or premise, where the premise either supports or attacks the claim or other premise. Evidence in the form of a premise is either support or attack. For intra-turn relations, we provided the following definitions and examples:

- **support:** a premise that supports the truth of the claim or premise. For example (premise in blue), *"Global stability is a myth grown out of a population that has gotten good at balancing acts. The supports necessary for real stability simply don't exist."*

- **attack:** a premise that attacks the truth of the claim or premise. For example (premises in red), *"Living and studying overseas is an irreplaceable experience. One will struggle with loneliness but those difficulties will turn into valuable experiences."* The example illustrates that the argument begins with a claim, and is followed by attacking one's own claim, and in turn attacking that premise (an example of the rhetorical move known as prolepsis or prebuttal).

In some cases, premises support the claim only if considered together (also known as "linked premises") and in such cases we instructed the annotators to label the two premises together as instantiating one support relation. Consider the argument *"Health insurance companies should naturally cover alternative medical treatments. Many general practioners (sic) offer these parallel treatments in parallel anyway and who would want to question their broad expertise?"*

Inter-turn relations connect the arguments of two participants in the discussion (agreement or disagreement). The argumentative components involved in inter-turn relations are claims, as the nature of dialogic argumentation is a difference in stance. We provided annotators with the following definitions and examples:

- **agreement:** a relation expressing an agreement or positive evaluation of another user's claim. For example, one user writes *"The only constant is change. Global stability is a myth grown out of a population that has gotten good at balancing acts. The supports necessary for real stability simply don't exist"* and another simply responds *"Good arguments."* Agreement, however, is not limited to expressions of agreement but may also be restatements or summarizations of the original post.

- **attack:** a relation expressing disagreement or a negative evaluation with the claim of another user. We also defined two types of attacks:

  - **rebuttal:** when a claim directly challenges the truth of the other claim. For example, User A may state *"I think some of the biggest threats to global stability comes from the political fringes."* and User B responds *"The only constant is change,"* disputing the idea that there even is global stability.

– **undercutter:** when a claim directly challenges the reasoning connecting the other claim to the premise that supports it. In this case, the full argument from User A (premise in blue) is *"I think some of the biggest threats to global stability comes from the political fringes. It has always been like that in the past."* User C responds directly to the premise, claiming *"What happened in the past has nothing to do with the present,"* disagreeing with the reasoning that the past is always a good predictor of current events.

- **partial agreement/attack:** where the response concedes part of the claim, stating two claims that express different stances. These claims are annotated together as instantiating one relation of partial agreement/attack, depending on the main stance. For example, User Z writes *"There are many things we used to consider extreme that are now fundamental principles of our society."* and User Y responds *"While this may be true, hasn't the killing of innocents (and murder or rape, or destruction of property) always been considered deeply immoral (outside of wars)?"*

4.1.3 Annotation Results

In total, the 112 discussion threads comprise 380 turns of dialogue for 2756 sentences. There were 2741 argumentative propositions out of which 1205 are claims and 1536 are premises, with an additional 799 non-argumentative propositions. At the sentence-level, 40.8% of sentences contain a claim and 52% contain a premise. 20% of sentences contain no annotations at all. In terms of claims,[4] 16.3% of sentences contain a rational evaluation, 12.6% contain an interpretation, and 8.6% contain an emotional evaluation, while only 3.1% contain agreement and 2.4% contain disagreement. For premises, 45.6% contain logos, 30% contain pathos, and only 4.3% contain ethos.

In terms of relations, 66% were in support, 26% attacking, and 8% partial. Overall, there are 7.06 sentences per post for our dataset, compared to 4.19 for the dataset of Morio and Fujita

[4] We took the majority vote among Turkers to determine the types of claims and premises.

[2018b]. This results in a large number of possible relations, as all pairs of argumentative components are candidates. The resulting dataset is hence very unbalanced (only 4.6% of 27254 possible pairs have a relation in the intra-turn case with only 3.2% of 26695 for inter-turn).

Claims and Premises We computed Inter-Annotator Agreement for claims and premises by requiring 3 of the expert annotators to annotate an overlapping subset of 2 threads. We compare annotations at the sentence level, similar to previous work [Stab and Gurevych, 2014a], as most sentences contain only 1 proposition, making this approximation reasonable. We compute IAA using Kripendorff's alpha [Krippendorff, 1970], obtaining 0.63 and 0.65, respectively. These scores are considered moderate agreement and are similar to the results on persuasive essays [Stab and Gurevych, 2014a]. Table 4.1 shows a case where two of the annotators disagreed. Both Annotator

Annotator	Annotation
1	[I would define patriotism quite simply as supporting one's country][**CLAIM**], [but not *necessarily* disparaging others][**CLAIM**]. [Therefore I would say that the antithesis of patriotism is not having a global perspective][**CLAIM**], but [is in fact selfishness][**CLAIM**].
2	[I would define patriotism quite simply as supporting one's country, but not *necessarily* disparaging others][**CLAIM**]. [Therefore I would say that the antithesis of patriotism is not having a global perspective][**PREMISE**], but [is in fact selfishness][**PREMISE**].

Table 4.1: An example where two annotators differ in their assessment of argumentative segments and labels.

1 and 2 agree that the definition in the first sentence consists of a claim. However, Annotator 1 splits the sentence into two claims based on the discourse connective *"but,"* whereas 2 has the correct interpretation, as the discourse connective is part of the definition and not a separate claim. Annotator 2 again labels the second sentence correctly, as the author of the post provides logical reasoning based on the definition and uses the explicit causal connective *"therefore."* Without the context of the first sentence, the second sentence could be interpreted as a claim due to the phrase *"I would say that."* This example illustrates the importance of providing context and the necessity of expert annotators due to the complexity of the task.

		Expert			Crowd		
Example	Premise	Logos	Pathos	Ethos	Logos	Pathos	Ethos
1	Someone who attempts to extract gain for themselves at harm to their overall country is acting unpatriotically.	x	x	-	x	x	-
2	A person doing that is probably an asshole, but that does not mean that silencing him is a proper reaction.	x	-	-	-	x	-
3	It seems that you can barely fathom that you *may* be wrong, that you are too intelligent to have to provide any adequate reasoning for yourself.	x	-	x	x	-	x

Table 4.2: Examples of premises and the corresponding expert and crowd judgments.

Semantic Types of Premises We also compute IAA for types of premises, comparing the majority vote of the Turkers to gold labels from our most expert annotator (based on highest average pair-wise IAA). As Kripendorff's alpha is calculated globally and compares each item directly between annotators, it is well-suited for handling the multi-label case here [Ravenscroft et al., 2016]. The resulting IAA was 0.73. This task was easier than the first stage as it requires only labeling a text segment, whereas the first stage requires both token-level annotation of text segments and labeling of the entire segment. We present examples of agreement and disagreement between gold labels and crowd-sourced labels in Table 4.2. In Example 1, both the expert and crowd annotators agree that this claim is both logos and pathos, as the author provides an example but with a latent emotional appeal to patriotism. In the second example, the expert considers this to be reasoning by example, but the annotators consider this to be an emotional appeal. These examples suggest that both expert annotators and crowd-workers often agree on the emotional connotations of words but differences arise due to latent interpretations. Finally, Example 3 shows that ethos is straightforward to identify, as appeals to authority consist of either establishing credibility or personal attacks.

L \ M	C_A	C_D	C_{EE}	C_{ER}	C_I
C_A	186	8	17	35	19
C_D	6	133	18	53	35
C_{EE}	21	35	424	187	112
C_{ER}	45	56	157	1150	220
C_I	23	45	105	205	459

Table 4.3: Confusion Matrix for Claims **L**: individual labels **M**: majority vote

Semantic Types of Claims Furthermore, we compute IAA for the types of claims, again comparing the majority vote to gold labels annotated by an expert linguist. The resulting IAA is 0.46, considered low agreement. This result is in line with those attested in similar experiments [Walker et al., 2012b]. We hypothesize that the low agreement for the types of claims may be due to the nature of the claims provided as the unit of annotation. According to the expert linguist annotator, some of the claims are complex sentences being formed by two propositions liable to two different types of claims, which may have led to confusion. In a sentence such as *"Your first paragraph is intriguing, and I definitely agree with it,"* for instance, the first proposition constitutes an emotional-evaluation, with the second an agreement. The choice of one of the two labels may, thus, give rise to divergent annotations.

To investigate the disagreement space in the annotation of types of claims, we present a confusion matrix in Table 4.3 between the majority vote and the label chosen by each of the 5 Turkers. The major disagreement is between the claim types "interpretation" (C_I) and "evaluation-rational" (C_{ER}), followed by the pairs "evaluation-emotional" (C_{EE})/ "evaluation-rational"(C_{ER}). While the label "disagreement" (C_D) also seems to be controversial, the scarcity of occurrences makes it less relevant for the analysis of the disagreement space. The higher consensus in the labeling of "agreement"(C_A) versus other types of evaluations can be explained looking at linguistic triggers: "agreement" is often signaled by unambiguous linguistic clues (*"I agree," "you are right," "yes"*).

In order to verify and explain difficulties encountered in deciding whether the claim is C_{ER} or C_I we compared the Turkers annotation with the gold annotations of an expert linguist annotator (Table 4.4). The trends in the disagreement space are the same as those noticed among Turkers.

Example	Claim	Expert	Crowd
1	Fair enough haha	Agreement	Agreement
2	Conservatives don't do this, it's extremists.	Disagreement	Evaluation-Rational
3	The problem isnt always bad parenting, though that can play a role, the problem is a black and white educational system	Interpretation	Evaluation-Rational
4	I don't think I'm better than the people I'd be denying citizenship	Evaluation-Emotional	Evaluation-Emotional
5	This is the best argument I've seen here	Evaluation-Rational	Evaluation-Rational

Table 4.4: Examples of claims and the corresponding expert and crowd judgments.

While expressions of agreement are linguistically diverse, they are often explicit and straightforward to identify, such as Example 1 in Table 4.4. On the other hand, while disagreement may be expressed explicitly (e.g. *"I don't think so"*), we find in our qualitative analysis that disagreement is most often expressed by re-phrasing the views of the other poster (e.g. Example 2), thus making it harder to identify. Our analysis also shows that Turkers tend to misclassify interpretation (C_I) as evaluation-rational (C_{ER}), as with Example 3. This is mainly due to a tendency of annotating claims as evaluations in the presence of a sentiment word regardless of the overall meaning of the proposition: the sentence *"The problem isnt always bad parenting, though that can play a role, the problem is a black and white educational system"* was annotated as an evaluation probably due to the axiological adjective *"bad."* However, the primary meaning is not that of providing a negative judgment, but that of providing an explanation for a state of affairs (problems encountered at school). Furthermore, the degree of rationality/emotions conveyed by a judgment is not always transparent given the semantics of the sentiment expressed, but may call for wider contextual features. Given a sentence such as *"I don't think I'm better than the people I'd be denying citizenship"* (Example 4) it is clear that what the speaker is expressing is a subjective evaluation, while in the sentence *"This is the best argument I have seen"* (Example 5) the type of evaluation at stake depends on the criteria at the basis of the judgment.

Relation Presence and Types Finally, we obtain moderate agreement for relation annotations, similar to other argumentative relation tasks on dialogues [Morio and Fujita, 2018b]. The Inter-Annotator Agreement (IAA) with Kripendorff's α is 0.61 for relation presence and 0.63 for relation types. This task is made easier by three factors. First, posters tend to provide evidence for claims and premises immediately due to the nature of discourse. For example, all annotators agreed that the claim *"Are we honestly expected to believe that no vacancies opened up at the inn for almost two weeks? "* was supported by the premise *"Because that would be the most unbelievable aspect of the entire story if you ask me."* Second, in inter-post relations, posters often quote the proposition to which they are responding (see Section 3.1.2). In one instance, an original poster states *"I believe smart phones are prevalent enough among phone users who want to use the Internet that it outweighs the "some people don't have smartphones" argument."* In the response, the poster directly quotes this claim and undercuts the claim with a new interpretative claim *"there are plenty of cheaper smartphones with less memory and a lesser processor which may chug a little bit on these websites."* Finally, in both inter-post and intra-post relations, posters often respond directly to the main claim. For the main claim *"Veganism and vegetarianism are not the best way to improve farm animal welfare"* the claim *"The chances of the entire world adopting a vegan/vegetarian diet are slim to none"* directly agrees with this main claim.

4.1.4 Analysis

In order to investigate what conceptual features are persuasive, we first observe correlations between types of argumentative components (premises and claims) as well as their position in the post. We then look at how different patterns are distributed in positive and/or negative threads.

Argumentative Components We present an analysis of correlations between types of claims and premises to better understand what types of relations are most common and provide a foundation for determining what types of relations are most effective. First, we check the presence of an ordering effect to determine what patterns recurrent models such as LSTMs and CRFs may be able to recognize. In the absence of a reliable model for argumentative components and their semantic types, neural models may still be able to identify the same patterns given the text. Second, we report the interaction of semantic types between premises and their previously occurring claims. Even if we assume we have a reliable model for identifying claims and premises and semantic types, we may not have a reliable model for argumentative relations. In this scenario, we may have a model (neural or otherwise) which models the interaction between predicted premises and their most recent predicted claim, which is a reasonable assumption, as premises often directly follow a claim. Finally, we report the results for gold claims, premises, semantic types, and their argumentative relations. In order to have a balanced dataset between winning and non-winning arguments, we use the same subset of 39 threads annotated in the first version of our corpus [Hidey et al., 2017] for all of our analysis.

We first report the results of the sequential transitions at the proposition level between types of claims (agreement, disagreement, rational evaluation, emotional evaluation, and interpretation) and premises (pathos, ethos, and logos, and their respective combinations). If the previous proposition is not labeled as claim or premise, we set the previous category to "None." If the sentence is the start of a post, we set the previous category to "BOP" (beginning of post). We also include transitions to the end of the post (EOP). We present results for the annotations from the AMT workers in Figure 4.4. The heatmap represents the transition matrix, normalized by the row count. The rows represent the label for the previous proposition and the columns represent the label for the current proposition.

We compute significance for individual cells using the chi-squared test for cells, computing a 2x2 contingency table. All results discussed have $p < 0.001$ after the Bonferroni correction, unless otherwise specified. Considering only claims at the beginning of the post, rational evaluations (23%), agreements (5%), and interpretations (13%) are more likely to appear at the start than in general. On the other end, premises expressing *pathos* are less likely to appear at the end of the post (only 7% of the time), while less surprisingly, unannotated sentences (farewell messages, for example) are more likely to appear at the end (20% of the time). As far as sequences of modes

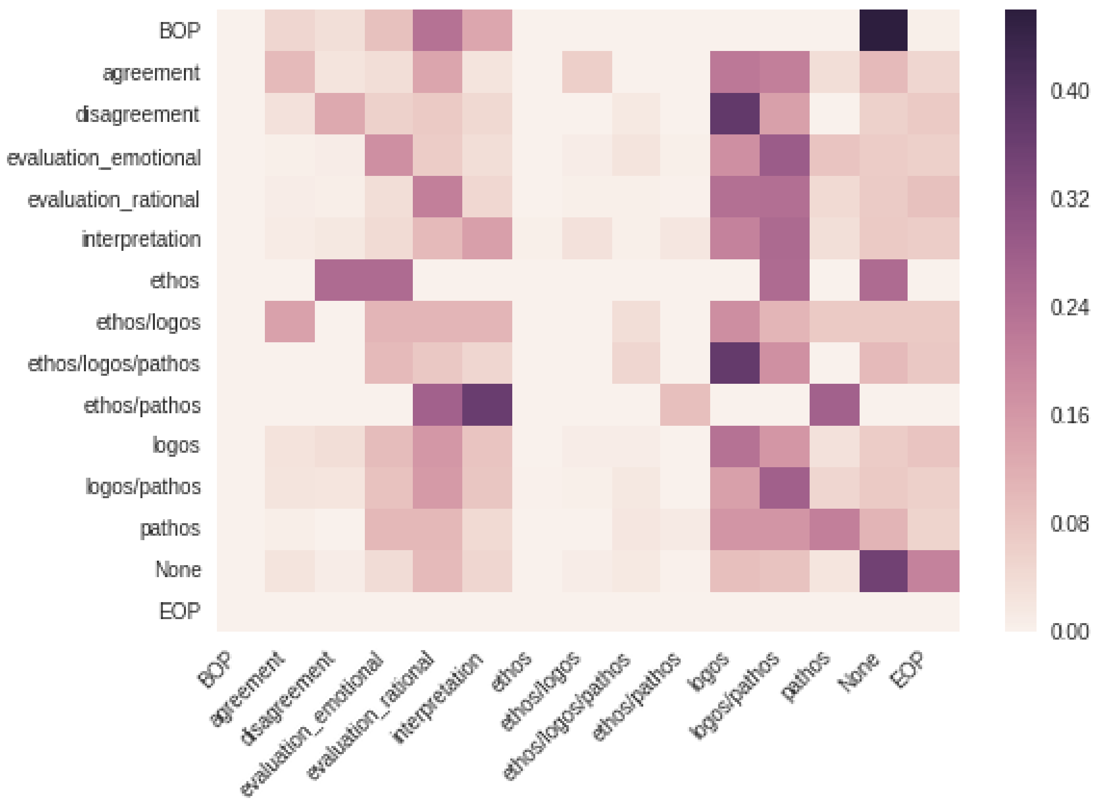


Figure 4.4: Transition Heatmap. Darker-shaded cells represent more likely transitions, e.g. that non-arguments (None) are likely to be at the beginning of the post (BOP).

of persuasion, arguments expressing *logos* or *pathos* are more likely to occur consecutively (for logos, 46% following logos and 48% following pathos and for pathos, 31% and 34% respectively) than in the overall distribution (37% *logos* and 24% *pathos*). Finally, *logos* is more likely to follow a rational evaluation (49% of the time) when compared to the overall distribution of *logos* and the same is true for emotional evaluations and pathos (39%).

For the second approach, we report the counts for the type of premise given the most recent claim type in the post. We assume here that the premise always attaches to the preceding claim, providing an approximation for this type of structure. We chose this heuristic since we observed that users tend first to express their view and then back it up with subsequent arguments to achieve a clear argument structure as advocated by Change My View submission rules. However, we acknowledge that premises may be positioned in front of a claim or refer anaphorically to a textually distant claim. We manually evaluated a sample of 100 premise-claim pairs: the correct pairs were identified 75% of the time. If the previous claim occurs either in the title or the previous post, we just indicate the previous claim to be "EOP." This scenario occurs when the original poster writes a premise that depends on the main claim or when a post responds directly to a claim in a preceding post. The heatmap in Figure 4.5 represents the conditional distribution of claims given the following premise. The rows represent the label for the claim and the columns represent the label for the subsequent premise, normalized by the counts of premises.

We again compute statistical significance with the chi-squared test, this time for pairs of premises and the immediately preceding claim. Premises classified as *pathos* are in support of rational evaluations 34% of the time that pathos occurs, while *logos* supports rational evaluations 38% of the time ($p < 0.05$) and *ethos* 28% of the time. Similarly, there is a slight preference ($p < 0.05$) for pathos to support evaluation-emotional claims, with 20% of pathos arguments supporting that type, 17% of logos arguments and 17% of ethos supporting it, respectively. Finally, authors demonstrate a preference for *logos* when addressing the claims of an author in the previous post ($p < 0.01$). The qualitative analysis of those cases reveals that when supporting rational evaluations, *pathos* arguments refer to situations that everyone could experience, as underlined by the use of the pronoun *you* in its impersonal use (e.g. *"If you don't break up, you are stuck with a person who doesn't value you enough to stay loyal. It's just a logical conclusion that breaking up is the right choice in most if not all situations."*)

For the third approach, we report the counts for the type of claim/premise given the claim/premise that it supports or attacks. Unlike the second approach, if the related claim occurs in a previous post, we know the semantic type given gold relations. We report both intra-post relations and inter-post relations. For intra-post relations, in our annotation scheme, a premise supports or attacks a claim or premise. In Figure 4.6, we report the type of premise and the type of claim/premise it supports or attacks. In other words, the heatmap visualizes the probability of the type of supported/attacked claim/premise given the type of supporting/attacking premise. For inter-post relations, the claim always agrees with or attacks a claim or premise from the previous poster or introduces a new topic ("None" in Figure 4.7). In Figure 4.7, we report the distribution of claims

Figure 4.5: Premise and Previous Claim Heatmap. Darker shades represent higher conditional probabilities for a claim type given a premise type.

given the claim or premise they respond to. In other words, the heatmap visualizes the probability of the type of supported/attacked claim/premise given the type of agreeing/attacking claim.

Using gold relations, we find that posters are more likely to use emotional claims to respond to emotional claims of the other poster ($p < 0.0001$) and less likely to respond to interpretative claims $p = 0.016$, similar to the analysis of previously occurring claims. Likewise, we find that *logos* premises are more likely to relate to interpretations ($p < 0.005$) and less likely to relate to emotional claims ($p < 0.005$), whereas pathos is less likely to relate to interpretations ($p = 0.011$). Interestingly, we find that pathos is more likely to relate to agreement claims ($p = 0.021$), suggesting that agreement has an emotional component. In terms of structure, posters are more likely to rebut the opposing claims rather than undercutting premises and to support claims directly with premises, forming a flat tree rather than a long chain of reasoning.

Semantic types and persuasive role To investigate whether certain types of claims/premises correlate with persuasive/non-persuasive messages, we conduct a preliminary analysis of the relationship between claims and premises in different contexts- in winning vs. non-winning arguments. We re-compute the transition matrix between premises (Figure 4.4) and the conditional claim/premise matrices (Figures 4.5, 4.7, and 4.6) separately for all winning arguments and non-

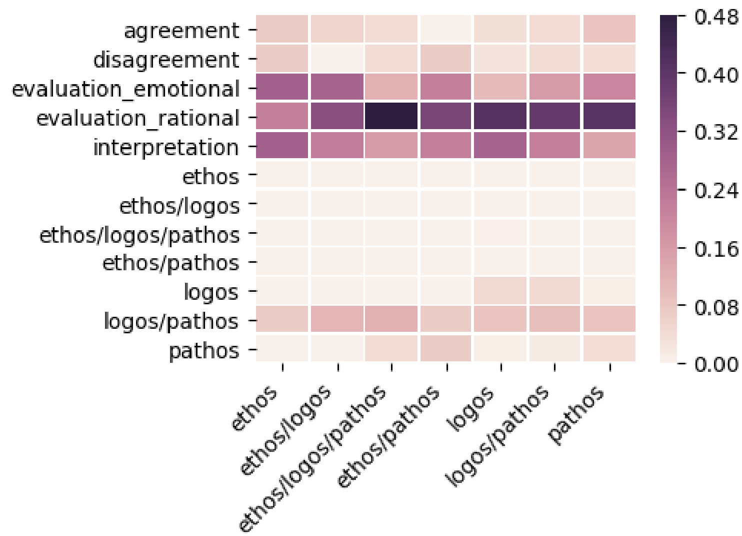


Figure 4.6: Premise Relation Heatmap. Darker-shaded cells represent higher conditional probabilities, e.g. that premises with ethos, logos, and pathos are more likely to support evaluation-rational claims.

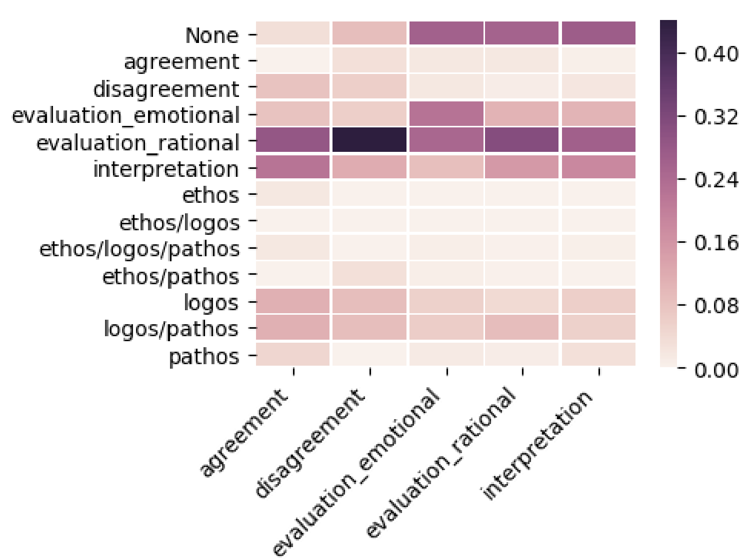


Figure 4.7: Claim Relation Heatmap. Darker-shaded cells represent higher conditional probabilities, e.g. that disagreement is more likely to occur in response to evaluation-rational claims.

winning arguments (i.e. by splitting the dataset according to whether the responding poster received a delta or not). As our goal is to understand whether certain patterns are more likely to be persuasive, we also consider only the components written by the author of the response and discard the posts from the original poster.

We compute statistical significance between the positive and negative label distributions and conditional and transition matrices using Pearson's chi-squared test of independence. As the chi-squared test considers the distribution of the data and does not require equal sample sizes,[5] this test is appropriate for significance. We again use the Yates correction for low frequencies. For the AMT annotations, we obtain a p-value of $p < 0.00001$ for all distributions: the unigram labels, the transition matrix, and the claim/premise matrices. For the gold annotations, the p-value of the overall label distribution is $p < 0.05$, but for the transition matrix the p-value is $p = 0.59$, likely due to the very low counts for some cells. However, the value for the prior claim/premise matrix and gold relation claim/premise matrices is $p < 0.001$, indicating significant differences even for this small dataset.

Next, similar to the analysis of the entire dataset, we compute significance for individual cells using the same chi-squared test. We first find that for the unigram distribution rational evaluations are *less* likely to be found in winning arguments with 9% of propositions in positive and 14% in negative ($p < 0.01$). When we consider the joint distribution of premise combinations, we find that *pathos* and *logos* are more likely to occur together in successful threads, with 23% and 17% respectively ($p < 0.01$).

For the transition distribution, compared to positive threads, negative threads show fewer agreements opening up the posts ($p < 0.05$). Agreeing with what was previously said by another speaker before expressing a possibly divergent opinion constitutes a traditional persuasive rhetorical strategy [Anscombre and Ducrot, 1983]. In a sentence such as *"I do agree that today's moderates are potentially tomorrow's conservatives. However this isn't about being just a bit conservative. It's about ..."*, the speaker concedes the previous user's point and then expresses a slightly contrasting

[5] Positive threads tend to be longer so they have more sentences and thus a higher number of claims and premises

point of view. In doing so, he exhibits his reasonableness and he avoids the face-threatening act [6] of disagreement. Moreover, positive threads are slightly more likely to show consecutive arguments of the same type (logos/logos; pathos/pathos) ($p < 0.01$), suggesting the hypothesis that conceptual coherence plays a role as persuasive strategy. The reasons provided by the original posters for awarding a Δ point frequently includes positive evaluations about the followed reasoning lines (e.g. *"Thanks for the brilliant and well thought out answer."*) For example, to support the logos premise *"Censorship does not eliminate the censored individual."* the poster write four consecutive logos-pathos premises: *"Indeed, they will work hard to evade the censor, carrying out their activities further from the public eye. This brings their arguments out of public discourse, where a solid counter-argument is more likely to be interjected. It also dooms any attempt to change their beliefs, as their opponents no longer understand those beliefs sufficiently to do this. Finally it gives them a true claim that they are being oppressed and targeted, which history suggests tends to strengthen a group."* In response to this specific argument, the OP states *"Good arguments"* and gives a delta *"for a well though out response."*

Finally, for gold relations, in winning arguments emotional claims are more often used to respond to *logos/pathos* premises of the other poster ($p = 0.048$), whereas rational claims are used more often in the same situation in non-winning arguments ($p < 0.01$), suggesting that undercutting a premise using the appropriate response is a valuable but underused strategy. Likewise, *logos/pathos* premises are used to support emotional claims in winning arguments ($p = 0.044$).

Examining premise/claim patterns qualitatively, it seems that positive threads generally feature more interpretations, especially based on arguments of the *logos* type, at the expense of the number of evaluations. This type of claim/premise pattern is likely to be perceived as less subjective. Evaluations, even when of the rational type, necessarily contain a subjective component in assessing the criteria to judge something as more or less positive or negative: the judgment *"networking is discriminatory"* during the hiring process would not, for instance, be shared by someone who considers social skills as a crucial quality for a job candidate. On the other hand, interpretations,

[6] `http://www.glottopedia.org/index.php/Face-threateningact`

when backed up by *logos*, encode states of affairs presented as intersubjective [Nuyts, 2012]. For instance, in the premise-claim pair *"American patriots have a general mentality against immigration. This is prominent in many ads and political champagnes, namely the slogan 'Creating jobs for americans' "* ads and political campaigns can be accessed by anyone. Since their goal is that of communicating a specific message to the public, the interpretation of their content promises to raise limited disagreement. This difference in degree of (inter)subjectivity is mirrored by the fact that evaluations, differently from interpretations, tend to be introduced by propositional attitude indicators at the first person singular (e.g. *"I think", "I find," "I point out"*) that put the speaker in a position of prominence as responsible for the truth of the asserted proposition. Moreover, evaluations are more frequently backed up by *pathos* arguments (e.g. the claim *"Enjoying the moment is possible, but doesn't make life have a point"* and the matching *pathos* premise *"For once I die, all memories and all point is gone"*.

4.1.5 Conclusions, Limitations, and Future Work

We have proposed an annotation scheme for the identification of persuasive conceptual features and empirically validated our approach using a three-stage process with both experts (for argumentative components and relations) and non-expert annotators (for semantic types). Compared to previous work in the same vein [Walker et al., 2012b, Habernal and Gurevych, 2017, Stab and Gurevych, 2014a], we unify argumentative structure, semantics, and persuasion in a single annotated corpus.

The annotation of argumentative components and relations achieves moderate agreement, in line with prior work [Stab and Gurevych, 2014a]. The same applies to the semantic types of premises, showing improvement with respect to previous attempts [Walker et al., 2012b, Habernal and Gurevych, 2017]. However, annotation of the semantic types of claims appears to be more difficult for non-experts due to the confusion between interpretations and rational evaluations. In future work, we plan to explore using expert annotators for this task, even though the role of expert annotators for argument structure is an additional limitation of our approach. Recent work has found that by using an iterative annotation approach [Miller et al., 2019] or by providing the appropriate context Lavee et al. [2019a], crowd workers can perform reasonably well at argumentative tasks. Thus we also plan to improve the guidelines to account for the difficulty of identifying the semantic types of claims.

While our annotation scheme resulted in moderate agreement for most tasks, we took a more coarse-grained approach than some theoretical work in argumentation mining, including argumentation schemes in dialogues [Reed and Walton, 2007]. Future work can examine how to annotate different models of semantics alongside our annotated data, allowing us to measure their persuasive impact. Furthermore, we annotated arguments at the token level, but we did not explicitly annotate the argumentative shell [Madnani et al., 2012], which would allow us to distinguish between the impact of content versus non-content words and phrases.

In order to understand the persuasive role of the semantic types of claims and premises, we observed combinations of argumentative components, their preferred position in the post, and their distribution in winning and non winning threads. For example, we observed that winning arguments tend to begin with agreement, tend to contain both pathos and logos, and tend to be semantically coherent. However, our analysis is only the beginning of the exploration of this data; future work could explore other aspects of persuasion. We expect that certain topics are more emotional or rational than others and winning arguments are generated accordingly. For example, moral issues may be more effective based on personal/emotional arguments while issues in science may require rational arguments. We also expect that the distribution of labels in the original post determines the effectiveness of a response, i.e. a post consisting mostly of emotional claims and pathos might require a similar response. Furthermore, additional fine-grained annotations for other argumentative strategies would allow one to explore additional effects.

4.2 Computational Models of Argumentation

Our analysis in Section 4.1 showed that certain patterns of usage (in terms of individual components and their interactions) are more effective at winning arguments. A first step, then, towards identifying persuasive arguments is the automatic identification of claims, premises, and their relations. As our goal is to build a system that can automatically detect winning patterns, we aim to make progress towards automation of argumentation mining in social media dialogues.

To this end, we introduce a large, novel dataset of distantly-labeled opinions from social media and demonstrate that by fine-tuning a language model on our relatively small dataset we improve performance on detecting claims and premises over a strong baseline [Chakrabarty et al., 2019b]. We also show that by framing the problem as claim detection at the sentence level, we improve over the previous state-of-the-art on multiple datasets [Chakrabarty et al., 2019b]. By modeling claims at the sentence level, we can make the comparison possible across datasets as different datasets use different units of annotation, thus giving us insight into the efficacy of our methods. Finally, we present a pipeline for argumentation mining. Given argument spans, we predict claims and premises and then predict the relations between them at both a micro (intra-post) level and a macro (inter-post) level [Chakrabarty et al., 2019c].

In Section 4.2.1, we first discuss how we leverage the social media acronyms and metadata discussed in Section 3.1.2 to **obtain distant-labeled data** and evaluate on our data from Section 4.1 and other datasets. We then present our approach to identifying argumentative components – claims and premises – by **leveraging recent advancements in transfer learning** [Howard and Ruder, 2018, Devlin et al., 2019] and fine-tuning a pre-trained language model on the distant-labeled data as an intermediate step (Section 4.2.2). Then we take a similar approach to identifying intra-post and inter-post relations– the distant-labeled data provides either discourse or dialogue context, respectively. In addition to these fine-tuning steps, we use additional methods for relation classification using discourse relations from Rhetorical Structure Theory (RST) and candidate selection using extractive summarization, combining these approaches in the full pipeline. Finally, in Section 4.2.3 we show that our approach improves over the previous state of the art and several strong baselines and in Section 4.2.4 we present a qualitative analysis to illustrate the benefits of our approach.

4.2.1 Data

We primarily conduct our experiments on the annotated data described in Section 4.1.3. To compare to previous work and evaluate the effectiveness of our methods at generalizing to out-of-domain data, we use additional datasets (web discourse and persuasive essays) from argumentation mining. Finally, we create two distantly-labeled datasets using the heuristics described in Section 3.1.2: the IMHO acronym and the quote feature. As the labeled datasets are small for modern deep learning approaches, we leverage the distant-labeled data from Reddit by using transfer learning techniques to fine-tune a language model on the appropriate context — micro-level (discourse context) for argumentative components and intra-turn relations and macro-level (dialogue context) for inter-turn relations.

Labeled Data We use four datasets to train and evaluate our models - our 112 annotated threads from **CMV** described in Section 4.1.3 and three additional datasets from prior work. As argumentation appears in both monologue and dialogue data, we choose an additional dataset from social media and two datasets created from student essays. Peldszus and Stede [2016a] created a corpus of German **microtexts** of controlled linguistic and rhetorical complexity. Each document includes a single argument and does not exceed five argumentative components. This corpus was translated to English, which we use for our experiments. The **persuasive essay** corpus [Stab and Gurevych, 2017a] includes 402 student essays. The scheme comprises major claims, claims, and premises at the clause level. This corpus has been used extensively in the argumentation mining community. The corpus from Habernal and Gurevych [2017] includes user-generated **web discourse** such as blog posts, or user comments annotated with claims and premises as well as backings, rebuttals and refutations. As with Daxenberger et al. [2017], when comparing methods across datasets, we model claim detection at the sentence level, as this is the only way to make all data sets compatible to each other. Table 4.5 gives an overview of the data.

	#Claims	#Sentences	%Claims
Microtext	112	449	24.94
Persuasive Essay	2108	7116	29.62
Web Discourse	211	3899	5.41
CMV	1206	3541	34.0

Table 4.5: Table showing number of claims and total number of sentences in the data sets along with the percentage of claims in them

Micro-level Context Data In order to leverage transfer learning methods, we need a large dataset with distant-labeled opinions and relation pairs. First, we use the method described in Section 3.1.2 to obtain data self-labeled for opinions using the acronym IM(H)O. We collect Reddit comments from December 2008 to August 2017 through the pushshift API,[7] resulting in 5,569,962 comments. Due in part to the size of the data, a diverse set of topics are represented, ranging from sports (e.g. *"**IMO**, Lakers are in big trouble next couple years"*) to treatment of animals (e.g. *"That's virtually the same as neglect right there **IMHO**."*) To use these examples for fine-tuning, we need only to remove the acronym (and any resulting unnecessary punctuation). We perform sentence and word tokenization using Spacy.[8] We then extract only the sentence containing IMO or IMHO and discarded the surrounding text. We refer to the resulting collection of comments as the **IMHO** dataset [Chakrabarty et al., 2019b].

As the sentence following a claim is often a premise (also supported by our research in Section 4.1.4), we can also use the following sentence (when present) to create distant-labeled premise and relation pairs. This assumption provides us with a dataset distant-labeled for *discourse context*, as the properties of claims and premises are inherently contextual and they are often linked by explicit or implicit discourse relations. The resulting dataset contains 4.6 million comments in total (as the following sentence is not always present and claims are not always supported by premises). We denote this dataset as **IMHO+context** [Chakrabarty et al., 2019c].

Macro-level Context Data While the IMHO data is useful for modeling *discourse* context from consecutive sentences from the same author, inter-turn relations are of a dialogic nature and would

[7]`http://pushshift.io`
[8]`https://spacy.io/`

benefit from models that consider that macro-level context. We take advantage of the quote feature of Reddit described in Section 3.1.2. Particularly in CMV, this feature is used to highlight exactly what part of someone's argument a particular user is targeting. In the example in Table 3.2 in Section 3.1.2, the response contains an exact quote of a claim in the original post. This assumption provides us with a distant-labeled dataset for *dialogue context*, which captures properties of social interaction including paraphrasing and explicit or implicit speech acts such as agreement or disagreement. We collect 95,406 threads from the full CMV subreddit between 2013-02-16 and 2018-09-05 and find pairs of posts where the quoted text in the response is an exact match for the original post (removing threads that overlap with the labeled data). This phenomenon occurs a minority of the time, but we obtain 19,413 threads. When the quote feature is used, posters often respond to multiple points in the original text, so for the 19,413 threads we obtain 97,636 pairs. As most language model fine-tuning is performed at the sentence level, we take the quoted text and the following sentence as our distant-labeled inter-post pairs. We refer to this dataset as **QR**, for quote-response pairs [Chakrabarty et al., 2019c].

4.2.2 Methods

Identifying argumentative components is a necessary precursor to predicting an argumentative relation. For intra-turn relations, a premise may support or attack a claim or a premise from the same post. Conversely, for inter-turn relations, a claim may agree with or attack a claim or a premise from a different post. We thus take a multi-stage approach, where we first identify the **"source"** of a relation, which is a premise for intra-turn relations and a claim for inter-turn relations. Then, we predict whether there exists a relation between the source and a **"target,"** which may be any claim or premise among a candidate set of propositions. As there may be intrinsic properties of targets that make them more likely to be responded to or supported, we consider an optional intermediate step of identifying candidate targets. Furthermore, because intra-turn relations require modeling discourse context and inter-turn relations require dialogue context, we model each type of relation as a separate process.

Thus, we model this process as a pipeline: perform three-way classification on claims, premises, and non-arguments and then predict if an outgoing relation exists from the source premise/claim to a target premise/claim. In predicting these relations, we consider all possible source-target pairs of premises and argumentative components within a single post (for intra-turn) and claims from one post and argumentative components from another post (for inter-turn). The set of source-target pairs may be further reduced using a candidate target selection method. Our full multi-stage approach is as follows:

1. **Source Identification:** We fine-tune models for identifying claims and premises using pre-trained language models and our distant-labeled datasets.

2. **(Optional) Candidate Target Selection:** We explore methods to identify likely targets of relations using extractive summarization.

3. **Relation Prediction:** We fine-tune models for predicting the relations between sources and targets again using pre-trained language models and our distant-labeled datasets. These models are ensembled with a relation classifier using discourse relations.

This pipeline is diagrammed in Figure 4.8. We first fine-tune a pre-trained language model [Devlin et al., 2019] on the appropriate discourse (IMHO) or dialogue (QR) context. Then we again fine-tune the model on one of three tasks: detecting claims and premises, detecting intra-post relations, and detecting inter-turn relations. For detecting relations, we ensemble the fine-tuned model with a classifier trained using discourse relation features obtained from an RST parser. Inter-turn relation prediction has the additional intermediate step of selecting candidate targets using an approach adapted from extractive summarization.

As we use the same fine-tuning approach for three different tasks, we first discuss the pre-trained language modeling approaches we explored, before discussing each of the three stages sequentially. Our labeled datasets are fairly small for deep learning methods, so we leverage recent advances in transfer learning for natural language processing. We investigate the use of pre-trained language models, which have had a number of recent successes by fine-tuning on the dataset of in-

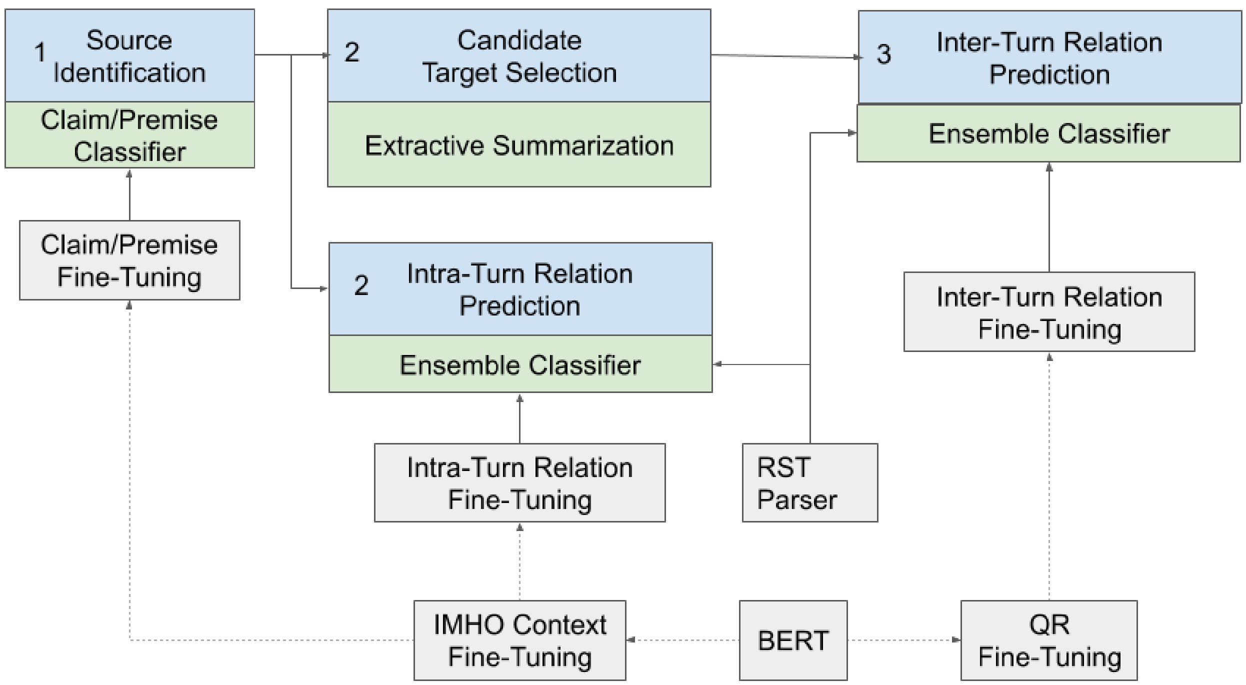


Figure 4.8: Our full pipeline for argumentative relation prediction. We first identify claims and premises as source candidates. Then we (optionally) identify target candidates using extractive summarization. Finally we predict the presence of a relation between a source-target pair.

terest [Howard and Ruder, 2018, Devlin et al., 2019]. Both the IMHO and IMHO+context datasets contain no negative examples, only labeled opinions (lacking non-claims). Furthermore, a self-labeled opinion may contain both a claim and a premise whereas an argumentative component consists of one or the other. We thus need a method of incorporating this dataset into an argumentative component detection model.

A comparison between two fine-tuning approaches is exhibited in Figure 4.9. The Universal Language Model Fine-Tuning method (ULMFiT) [Howard and Ruder, 2018] consists of the following steps: a) General-domain language model pre-training b) Task-specific language model fine-tuning and c) Task-specific classifier fine-tuning. In step (a), a stacked LSTM [Hochreiter and Schmidhuber, 1997, Graves et al., 2013] language model is trained on Wikitext-103 [Merity et al., 2017] consisting of 28,595 preprocessed Wikipedia articles and 103 million words capturing general properties of language. In step (b) the language model is trained on the task-specific dataset to capture domain-specific sequences of words and in step (c) a classifier is trained on the target task, fine-tuning the pre-trained LSTM but with an additional layer for class prediction. On the other hand, the Bidirectional Encoder Representations from Transformers (BERT) method involves step (a) general-domain language model pre-training and (c) task-specific classifier fine-tuning. In step (a), the BERT model is initially trained with a multi-task objective (masked language modeling and next-sentence prediction using a transformer model [Vaswani et al., 2017]) over a 3.3 billion word English corpus. In the standard use of (c), given a pre-trained BERT model, the model can be used for transfer learning by fine-tuning on a domain-specific corpus using a supervised learning objective. Unlike ULMFiT, there is no step (b) in the standard usage.

Inspired by step (b) of ULMFiT, we introduce an intermediate fine-tuning stage using our distant-labeled datasets. While ULMFiT trains a language model directly on the task-specific dataset, this approach is limited by the size of the dataset and thus the model is unable to learn complex interactions between words. We hypothesize that by using distant-labeled datasets for discourse and dialogue context as appropriate, the fine-tuned language models are better attuned to discriminative patterns for each type of data. Our approach is outlined in Figure 4.10. In step (b),

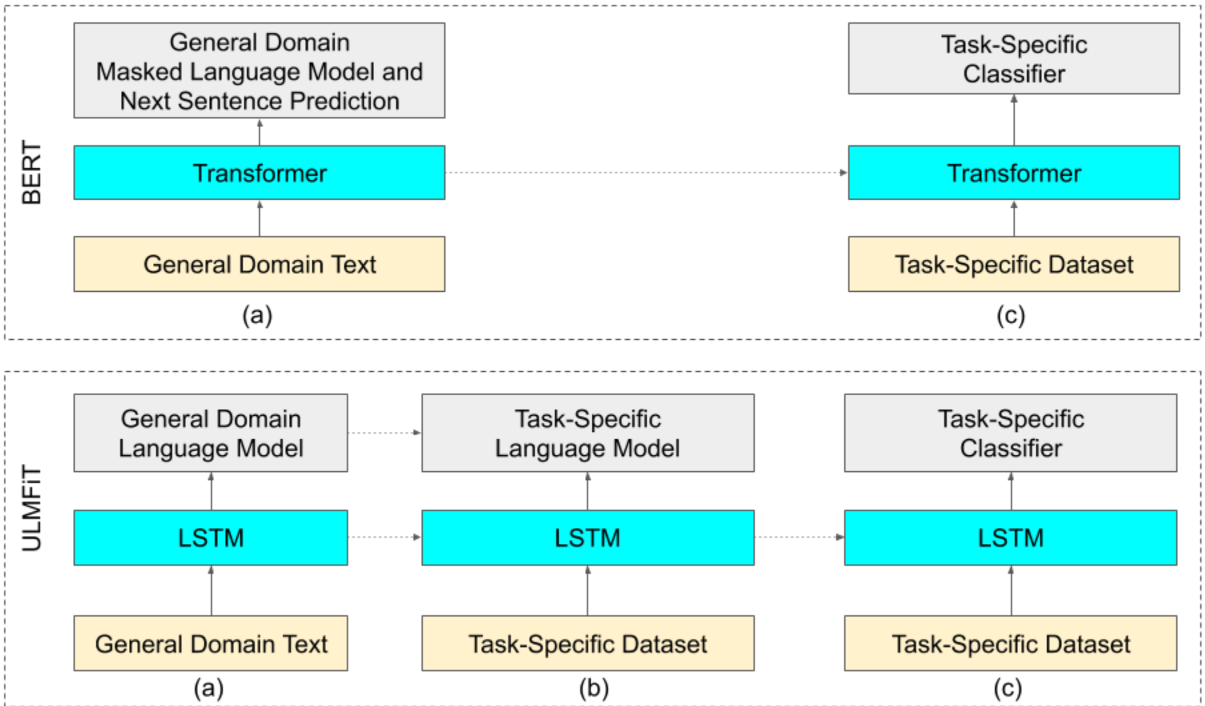


Figure 4.9: The standard BERT and ULMFiT training pipelines. The standard BERT usage does not include step (b) and instead directly tunes the classifier on the task data, as in step (c). The standard ULMFiT usage involves an intermediate language model fine-tuning step (b) on the same dataset used to train the classifier in step (c). Dashed arrows indicate that the parameters from the previous stage were used to initialize the next stage.

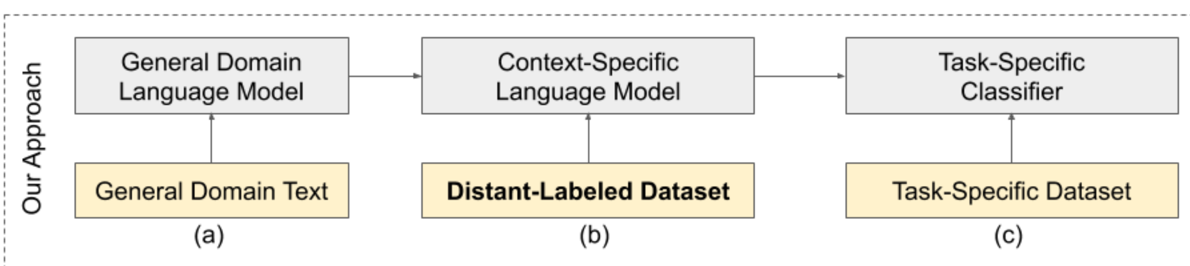


Figure 4.10: Our training pipeline. Unlike the standard ULMFiT or BERT usage, our step (b) involves fine-tuning on a distant-labeled dataset, rather than the dataset for classification in step (c). In our use case, we fine-tune on either IMHO/IMHO+context for argumentative component and intra-post relation prediction or QR for inter-post relation prediction.

we fine-tune a language model for the appropriate context on our distant-labeled datasets before fine-tuning a classifier in step (c) on one of our three tasks. As our approach is generally agnostic to the type of language model, we evaluate on both ULMFiT and BERT.

As ULMFiT has shown good performance on sentence-level tasks, we use this model for our claim detection experiments across datasets [Chakrabarty et al., 2019b]. On the other hand, BERT has been shown to work on both sentence prediction tasks such as sentiment detection and relation prediction tasks such as natural language inference [Bowman et al., 2015]. Accordingly, we use this model for our experiments on both relation prediction and argumentative component prediction [Chakrabarty et al., 2019c], as claims and premises are necessarily contextual. Furthermore, as the ULMFiT language model was pre-trained on single sentences, we fine-tune the model only on the IMHO dataset. In contrast, BERT was pre-trained on pairs of sentences so we fine-tune on IMHO+context.

An example of the intermediate fine-tuning on the IMHO+context dataset is presented in Figure 4.11. The original distant-labeled opinion is *"IMHO, Calorie-counting is a crock what you have to look at is how wholesome are the foods you are eating."* and the following sentence is *"Refined sugar is worse than just empty calories - I believe your body uses a lot."* As part of the masked language model fine-tuning, BERT is trained to recover missing words given the context.[9] Simultaneously, the model is trained to predict whether the second sentence follows the first sentence, compared to a randomly sampled sentence. This approach allows the model to learn interactions between words such as *"calorie"* and *"sugar,"* which are likely to occur in a discourse context. Similarly, an example of the intermediate fine-tuning on the QR dataset is presented in Figure 4.12. The sentence in one comment is *"A politician's sex has zero bearing on their ability to govern or craft effective legislation."* and the sentence in the response is *"Nobody is saying that women are better politicians than men, and thus, more female representation is inherently better for our political system."*. This approach allows the model to learn interactions between phrases such as *"sex"* and *"women"* which are likely to indicate related context and discriminative phrases

[9] In the ULMFiT case, the language model is trained to recover the word given only the previous words.

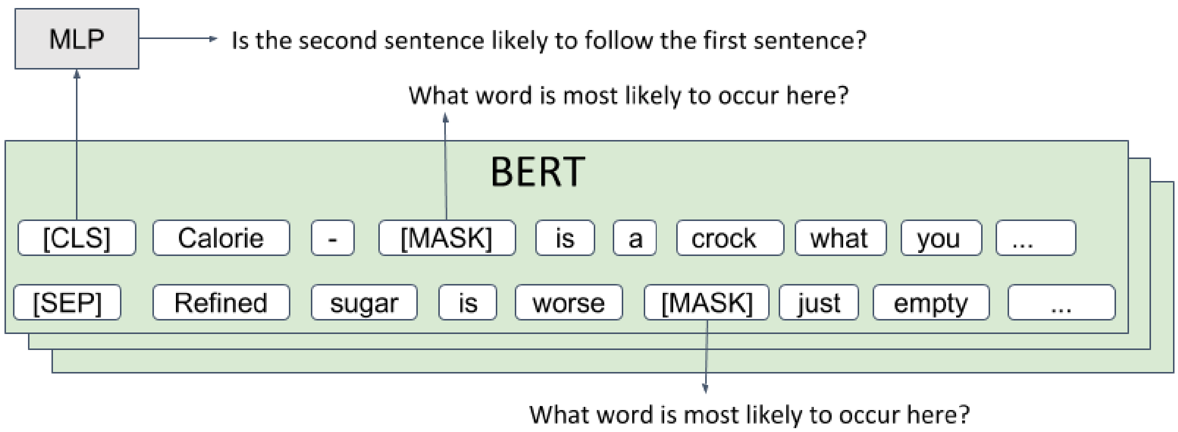


Figure 4.11: A sample training instance for IMHO+context fine-tuning.

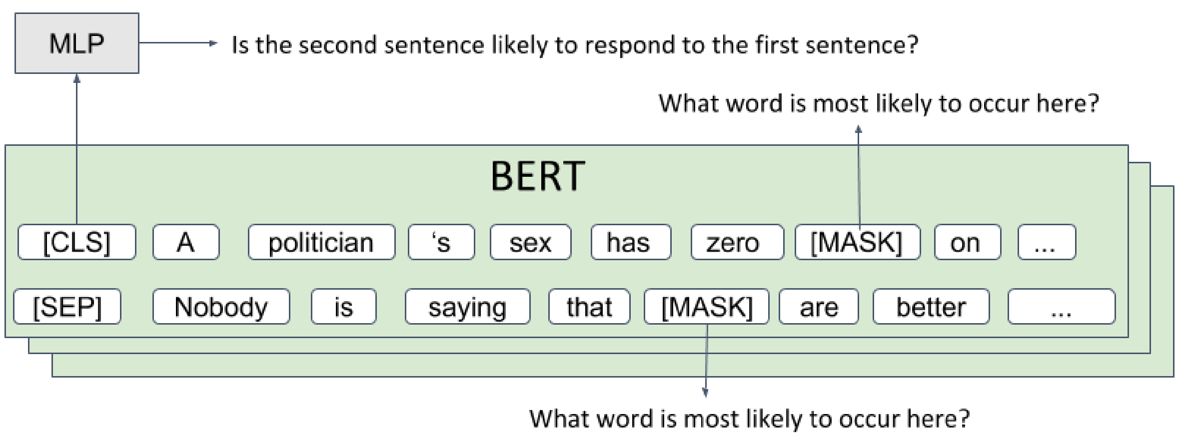


Figure 4.12: A sample training instance for QR fine-tuning.

such as *"nobody is saying that,"* which are likely to occur in a dialogue context.

The goal of ULMFiT and BERT is to allow training on small datasets of only a few hundred examples, but our experiments will show that fine-tuning the language model on opinionated claims improves over only task-specific LM fine-tuning. We hypothesize that this novel use of ULMFiT and BERT will help because the distant-labeled data is structured such that the sentence and next sentence (when present) will encourage the model to learn discriminative lexical features from a single claim or interactive features from a pair of sentences that improve performance on detecting argumentative components.

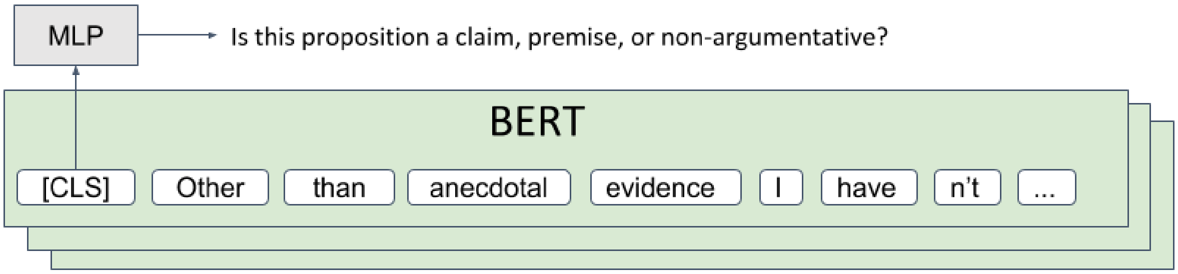


Figure 4.13: An example training instance for argumentative component classification.

Source Identification In the first stage of our pipeline from Figure 4.8, we identify claims and premises, as their possible usage as a source is dependent on the intra-post or inter-post context. Given the fine-tuned language model on the IMHO or IMHO+context dataset, we fine-tune a classifier to predict argumentative components. For ULMFiT, we fine-tune a classifier to make a binary prediction of whether a sentence contains a claim. For BERT, we fine-tune a classifier to perform three-way classification on claims, premises, and non-arguments. An example training instance for BERT is provided in Figure 4.13. In this example, the proposition is *"Other than anecdotal evidence, I haven't seen anything to support this claim."* and is labeled as a claim.

(Optional) Candidate Target Selection Next, for inter-turn relations, we take additional steps to reduce the number of invalid relation pairs. Predicting an argumentative relation is made more difficult by the fact that we need to consider all possible relation pairs. However, some argumentative components may contain linguistic properties that allow us to predict when they are targets even without the full relation pair. Thus, if we can predict the targets with high recall, we are likely to increase precision as we can reduce the number of false positives. Our candidate selection component, which identifies potential targets (as shown in Figure 4.8), consists of two sub-components: an **extractive summarizer** and a **source-target constraint**.

First, we use the **QR** data to train a model to identify candidate targets using techniques from **extractive summarization**, with the idea that targets may be salient sentences or propositions. We treat the quoted sentences as gold labels, resulting in 19,413 pairs of document (post) and

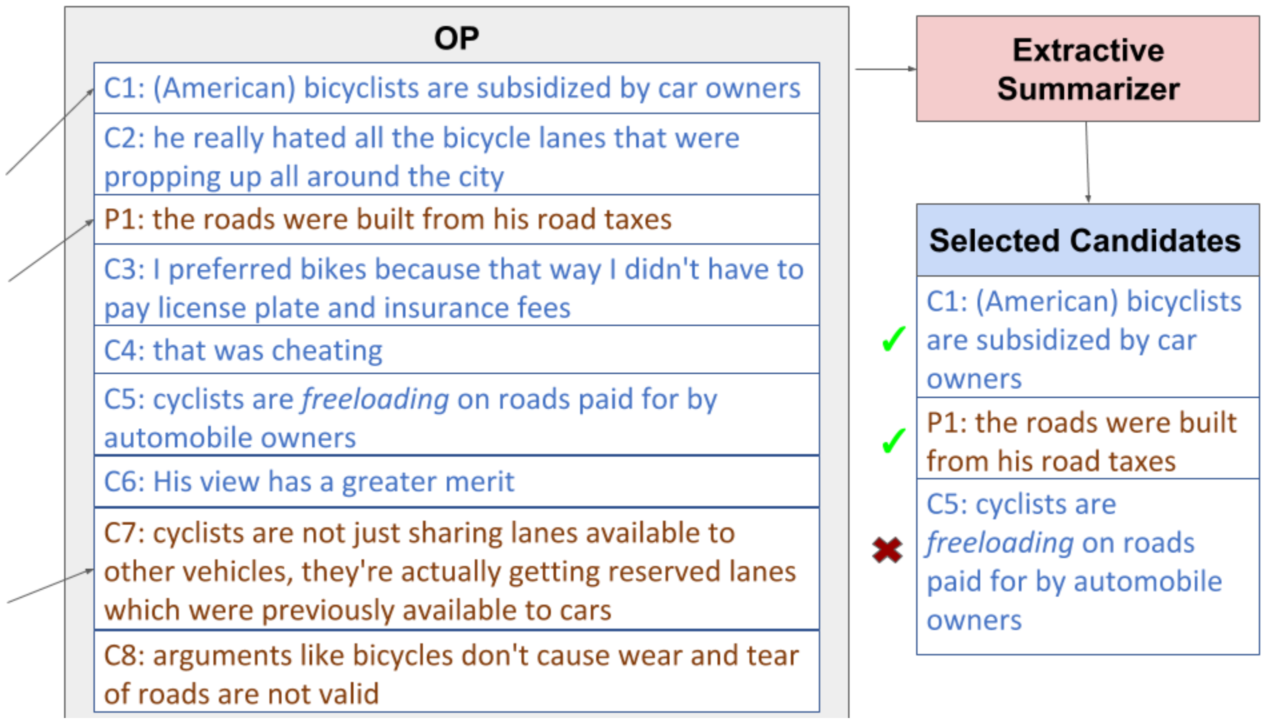


Figure 4.14: An example original post (OP) and the predicted targets. The incoming arrows to OP are the true targets, i.e. the propositions that were responded to.

summary (quoted sentences). For the example provided in Table 3.2, this would result in one sentence included in the summary. Thus, for a candidate source-target pair A → B, where B is the quoted sentence in Table 3.2, if B is not extracted by the summarization model we predict that there is no relation between A and B.

We use a state-of-the-art extractive summarization approach [Liu, 2019] for extracting the targets. The authors obtain sentence representations from BERT [Devlin et al., 2019], and build several summarization specific layers stacked on top of the BERT outputs, to capture document-level features for extracting summaries. We select the best summarization model on a held out subset using recall at the top K sentences. An example of target selection via extractive summarization is shown in Figure 4.14. In this example from Change My View, the response replies to C1, P1, and C7, and if $K = 3$, we would ideally select these three propositions as targets. However, the predicted targets are C1, P1, and C5. We hypothesize that the summarization model can learn to discriminate between generic claims such as C6 (*"His view has a greater merit"*), which are unlikely to be responded to, and salient claims and premises.

Second, in addition to summarization, we take advantage of a **dataset-specific constraint**: a target cannot also be a source unless it is related to the main claim. In other words, if B is a predicted target in A → B, we predict that there is no relation for B → C except when C is a main claim. In the CMV data, the main claim is always the title of the original Reddit post, so it is trivial to identify.

Relation Prediction In the final stage, we predict the *presence*[10] of a relation between a source and target pair. Given the predicted argumentative components as sources and selected targets, we consider all possible source-target pairs of premises and argumentative components within a single post (for intra-turn) and claims from one post and argumentative components from another post (for inter-turn). We then make a binary prediction of whether a relation is present between a pair of propositions. Our relation prediction module consists of two submodules: **fine-tuning** on the appropriate micro- and macro-level context for the respective intra- and inter-turn relations, and ensembling these models with a classifier trained with features derived from **RST relations**.

For intra-turn relation prediction we use the same fine-tuned BERT model on **IMHO+context** that we used for argument component classification, as premises often immediately follow claims so this task is a noisy analogue to the task of interest.[11] We then **fine-tune on the intra-turn relation prediction task** on all possible pairs *within* a post, using the labeled relations in the CMV data. For inter-turn relation prediction, rather than use **IMHO+context**, which consists of consecutive sentences from the same author, we use our model fine-tuned on the **QR** dataset, where the dialogue context more closely represents our labeled inter-post relations. Then, we **fine-tune on inter-turn relation prediction** using all possible pairs *across* two posts as training. An example of intra-turn relation prediction is presented in Figure 4.15. This example contains the claim *"How would you even quantify that?"*, which takes the form of a rhetorical question in a

[10]As we found that most intra-turn relations were in support and inter-turn relations were attacking, due to the dialogic nature of the data for our experiments we only predicted whether a relation was present and not the relation type.

[11]We do not use ULMFiT fine-tuning for relation prediction, as ULMFiT was pre-trained on single sentences and is thus not suitable for a sentence pair task.

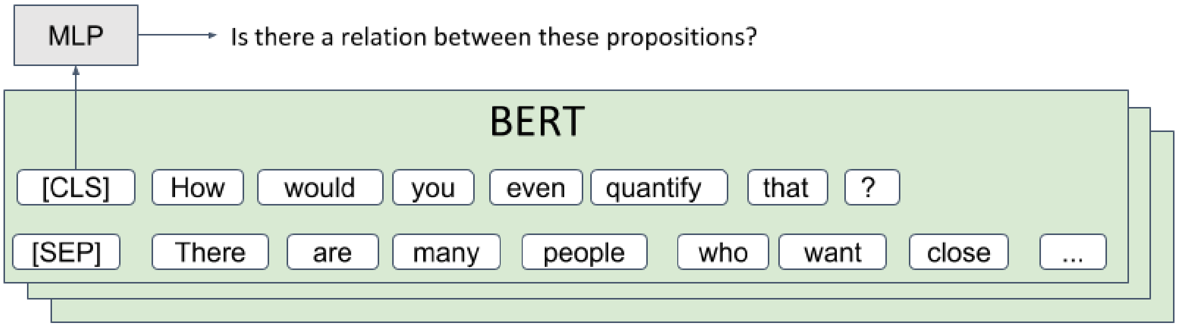


Figure 4.15: An example of an intra-post relation.

dialogue context, and the supporting premise *"There are many people who want close relationships without romance."* Ideally, the model would learn that *"quantify"* and *"many people"* would be likely to co-occur in a discourse context.

The second submodule consists of a classifier trained with features from **RST relations**. Rhetorical Structure Theory was originally developed to offer an explanation of the coherence of texts. Musi et al. [2018] and, more recently Hewett et al. [2019], showed that discourse relations from RST often correlate with argumentative relations. We thus derive features from RST trees and train a classifier using these features to predict an argumentative relation. To extract features from a pair of argumentative components, we first concatenate the two components so that they form a single text input. We then use a state-of-the-art RST discourse parser [Ji and Eisenstein, 2014][12] to create parse trees and take the predicted discourse relation at the root of the parse tree as a categorical feature in a classifier. For example, given the claim and premise *"If existence from your perspective lies solely on your consciousness, after you die it doesn't matter what you left,"* the predicted RST parse tree is illustrated in Figure 4.2.2. The predicted relation at the root node is a conditional statement, which we might expect to be indicative of an argumentative relation. There are 28 unique discourse relations predicted in the data, including *Circumstance*, *Purpose*, and *Antithesis*. We use a one-hot encoding of these relations as features and train an XGBoost Classifier [Chen and Guestrin, 2016] to predict whether an argument relation exists. This classifier

[12]We use Wang et al. [2018] for segmentation of text into elementary discourse units as they obtain the best results using ELMO [Peters et al., 2018].

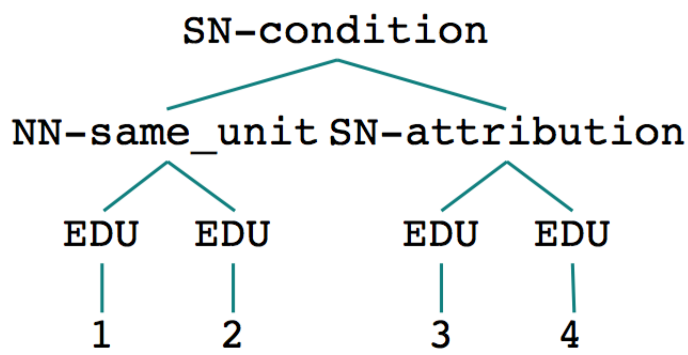


Figure 4.16: RST parse tree obtained from the sentence “If existence from your perspective lies solely on your consciousness, after you die it doesn’t matter what you left.”

with discourse relations, as indicated in Figure 4.8, is then ensembled with our predictions from the BERT classifier by predicting a relation if either one of the classifiers predicts a relation.

4.2.3 Experiments and Results

We train and evaluate our approaches on the labeled data from Section 4.2.1. We first conduct our **out-of-domain** experiments, demonstrating how fine-tuning on the IMHO comments generalizes to other datasets such as persuasive essays. Then, we present results on **argumentative component prediction**, comparing to state-of-the-art approaches to the task and other strong baselines. Finally, we present the results of the **full pipeline** compared to prior work on the same task.

Out-of-Domain Results We first compare our models and data to previous work on other datasets to show the effect of the IMHO fine-tuning. As each dataset has a different label scheme (i.e. some datasets have “warrant” labels, whereas we only have claim and premise) and unit of annotation (i.e. our dataset uses propositions whereas others use clauses or sentences), we frame the task as a claim detection task, where the goal is to predict whether a sentence contains a claim. Table 4.6 show the results on our CMV dataset, along with the other three described in Section 4.2.1. We compare to two baselines. The numbers in the **CNN** column are taken directly from the results of the deep learning experiments mentioned in the work of [Daxenberger et al., 2017]. Their deep

	Metric	CNN		Task-Specific LM Fine-Tuning		IMHO LM Fine-Tuning	
		Claim	Macro	Claim	Macro	Claim	Macro
Web Discourse	P	50.0	72.5	50.0	72.5	**54.0**	**75.9**
	R	20.4	59.2	20.0	59.8	**24.0**	**61.7**
	F	28.9	62.6	28.5	62.7	**33.3**	**65.2**
Micro-text	P	66.5	79.0	66.2	78.5	**71.0**	**80.9**
	R	68.2	78.5	68.0	77.8	**71.8**	**81.4**
	F	67.3	78.6	67.0	78.1	**71.2**	**81.1**
Persuasive Essay	P	60.9	73.2	62.3	73.2	**62.6**	**74.4**
	R	61.2	74.0	65.8	75.1	**66.0**	**75.0**
	F	61.1	73.6	64.0	74.1	**64.3**	**74.8**
CMV	P	54.0	65.1	55.0	68.0	**55.7**	**69.5**
	R	53.0	62.5	59.0	65.0	**60.0**	**65.3**
	F	53.5	63.8	57.0	66.4	**57.8**	**67.3**

Table 4.6: Table showing the results on four data sets. Each cell contains the Precision (P), Recall (R) and F-score (F) for Claims as well as the Macro Precision, Recall and F-score for the binary classification.

learning experiments consisted of 4 different models: a) bidirectional LSTM b) LSTM c) CNN initialized with random word embeddings and d) CNN initialized with word2vec. In their experiments for MT and PE, a CNN initialized with random word embeddings gave the best results and for WD a CNN with word2vec gave the best results. As CMV is a new data set we experimented with all four models and obtained the best result using a CNN with random initialization. The **Task-Specific LM Fine-Tuning** column contains the results obtained by fine-tuning the ULMFiT language model on each respective dataset while the **IMHO LM Fine-Tuning** column contains the results from fine-tuning the ULMFiT language model on IMHO.

The experiments were carried out in a 10-fold cross-validation setup with fixed splits into training and test data and the F1 scores are averaged over each of the folds. Each model was run 10 times to account for variance and the results reported in the table are an average of 10 runs. For our CNN experiments, we use the same hyper-parameters as Daxenberger et al. [2017]. For fine-tuning, we use the same hyper-parameters as Howard and Ruder [2018] except for a batch size of 32 for MT and 64 for the remaining data sets. The learning rate for classifier fine-tuning is set to 0.0001. We train our classifier for 5 epochs on each data set.

Method	C	P	NA
Argumentative Features + EWE	56.0	65.9	69.6
Joint Pointer Network	54.2	68.5	73.2
ULMFiT IMHO Fine-Tuning	57.8	70.8	70.5
BERT	62.0	72.2	71.3
BERT IMHO+Context Fine-Tuning	**67.1**	**72.5**	**75.7**

Table 4.7: F-scores for 3-way Classification: Claim (C), Premise (P), Non-Argument (NA)

We obtain statistically significant results ($p < 0.05$ with a chi-squared test) over all CNN models trained only on the task-specific datasets. We also find that for all models, IMHO LM Fine-Tuning performs better than Task-Specific LM Fine-Tuning, and is significantly better for the MT and WD datasets (which both contain very few claims). For the MT and WD datasets, Task-Specific LM Fine-Tuning actually performs worse than the CNN models.

Argumentative Component Results For baseline experiments on argumentative component classification we implement a model using the custom **argumentative features** of Stab and Gurevych [2017a]: lexical (unigrams), structural (token statistics and position), indicator (*I, me, my*), syntactic (POS, modal verbs), discourse relation (PDTB), and word embedding features. As shown in Section 4.1.4, emotional appeal or pathos is strongly correlated with persuasion and appears in premises [Hidey et al., 2017]. This motivated us to augment the work of Stab and Gurevych [2017a] with emotion embeddings [Agrawal et al., 2018] which capture emotion-enriched word representations and show improved performance over generic embeddings (denoted in the table as **EWE**).

We also compare our results to several neural models - a model using **joint pointer** networks [Morio and Fujita, 2018b] that was previously evaluated on Japanese dialogue, our model using **ULMFiT fine-tuning on IMHO** [Chakrabarty et al., 2019b], and a **BERT** baseline [Devlin et al., 2019] using only the pre-trained model without our additional fine-tuning step. We present the results of these models, along with our approach using **BERT IMHO+Context fine-tuning** in Table 4.7.

We set aside 10% of the data for testing, using the rest for training and validation. We use the implementation of BERT provided by Huggingface,[13] version 0.5. We fine-tune the language model on IMHO+context using the default hyper-parameters for 2 epochs, respectively. We fine-tune with a batch size of 128 and a learning rate of $2e-5$, training for 10 epochs.[14] For our baseline models, we use the XGBoost library[15] trained with the default settings.

Table 4.7 shows that our best model gains statistically significant improvement over all the other models ($p < 0.001$ with a chi-squared test). These results show that fine-tuning on the appropriate context is key. Furthermore, to compare directly to our work using ULMFiT [Chakrabarty et al., 2019b], we also test our model on the binary claim detection task and obtain a Claim F-Score of 70.0 with fine-tuned BERT, which is a 5-point improvement in F-score over pre-trained BERT and a 12-point improvement over our fine-tuned ULMFiT, suggesting that fine-tuning on context helps, although a more powerful model helps as well.

Full Pipeline Results We also report results on the full pipeline task using the CMV data from Section 4.1.3, where we first predict claims and premises and then argumentative relations. We compare a number of strong baselines against our system described in Section 4.2.2 for both intra-turn and inter-turn relations.

For our baseline experiments, we consider prior work in macro-level argument mining. Menini et al. [2018] predict argumentative relations between entire political speeches from different speakers, which is similar to our dialogues. We re-implement their model using their **argument relation features** (lexical overlap, negation, argument entailment, and argument sentiment, among others). As with component classification, we also compare to neural models for relation prediction - the **joint pointer network** architecture [Morio and Fujita, 2018b] and the pre-trained **BERT** [Devlin et al., 2019] baseline.

As the majority of component pairs contain no relation, we could obtain high accuracy by predicting that all pairs have no relation. Instead, we want to measure our performance on relations, so we also include an **"all-relation"** baseline, where we always predict that there is a relation

[13] `https://github.com/huggingface/pytorch-pretrained-BERT`
[14] These hyper-parameters were chosen based on the results on the validation set.
[15] `https://xgboost.readthedocs.io/en/latest/parameter.html`

Method	Precision		Recall		F-Score	
	Gold	Pred	Gold	Pred	Gold	Pred
All Relations	5.0	-	100.0	-	9.0	-
Argument Relation Features	7.0	5.9	82.0	80.0	13.0	11.0
Joint Pointer Network	10.0	-	48.8	-	16.6	-
BERT	12.0	11.0	67.0	60.0	20.3	18.5
Our Intra-Turn System	**16.7**	**15.5**	73.0	70.2	**27.2**	**25.4**

Table 4.8: Results for Intra-turn Relation Prediction with Gold and Predicted Premises

between two components, to indicate the difficulty of modeling such an imbalanced data set. In the test data, for intra-turn relations there are 2264 relation pairs, of which only 174 have a relation, and for inter-turn relations there are 120 relation pairs, compared to 2381 pairs with no relation.

As described in Section 4.2.2, for intra-turn relations, the source is constrained to be a premise whereas for intra-turn, it is constrained to be a claim. We thus provide experiments using both gold claims/premises and predicted ones. For intra-turn relation prediction, **Our Intra-Turn System** includes the IMHO+context fine-tuned BERT model ensembled with the RST classifier, as described in 4.2.2. **Our Inter-Turn System**, on the other hand, uses the QR fine-tuned BERT model, along with the RST classifier and candidate target selection components consisting of the extractive summarizer and source-target constraint.

As with argument component prediction, we set aside 10% of the data for testing, using the rest for training and validation. We use the same software and hyper-parameters for fine-tuning on IMHO+context and training the feature-based baseline models. We fine-tune BERT on QR using the default hyper-parameters for 3 epochs. We fine-tune with a batch size of 128 and a learning rate of $2e-5$ for intra-turn relation prediction and inter-turn relation prediction, training for 8 and 7 epochs, respectively. Using our extractive summarizer, we found that we obtained the best target recall of 62.7 at $K = 5$ (the number of targets to select).

We report the results of our binary classification task on **intra-turn relations** in Table 4.8 in terms of precision, recall and F-score for the "true" class, i.e., when a relation is present. We report results given both gold premises and predicted premises (using our best model from Table 4.7). As relation prediction is a difficult task, we obtain comparable performance to previous

Method	Precision		Recall		F-Score	
	Gold	Pred	Gold	Pred	Gold	Pred
All Relations	5.0	-	100.0	-	9.0	-
Argument Relation Features	5.9	4.8	82.0	80.0	11.0	9.0
Joint Pointer Network	7.6	-	40.0	-	12.7	-
BERT	8.8	7.9	76.0	70.0	15.8	14.1
Our Inter-Turn System	**18.9**	**17.5**	79.0	74.0	**30.5**	**28.3**

Table 4.9: Results for Inter-Turn Relation Prediction with Gold and Predicted Claims

Method	Precision		Recall		F-Score	
	Gold	Pred	Gold	Pred	Gold	Pred
Argument Relation Features	7.0	5.9	82.0	80.0	13.0	11.0
Argument Relation Features + RST Features	7.4	6.1	83.0	81.0	13.7	11.4
RST Features	6.3	5.7	79.5	77.0	11.8	10.6
IMHO+Context Fine-Tuned BERT	14.3	13.2	69.0	65.0	23.7	21.8
Our Intra-Turn System	**16.7**	**15.5**	**73.0**	**70.2**	**27.2**	**25.4**

Table 4.10: Ablation Experiments for Intra-turn Relation Prediction with Gold and Predicted Premises

work on relation prediction in other argumentative datasets [Niculae et al., 2017, Morio and Fujita, 2018b]. Our best results are obtained from ensembling the RST classifier with BERT fine-tuned on IMHO+context, for statistically significant ($p < 0.001$) improvement over all other models.

As with intra-turn relations, we report F-score on the “true” class in Table 4.9 for **inter-turn relations** using both gold and predicted claims. We again obtain statistically significant ($p < 0.001$) improvement over all baselines. Our best results are obtained by fine-tuning the BERT model on the appropriate context (in this case the QR data) and ensembling the predictions with the RST classifier along with our candidate target selection approaches.

System Component Ablation To better understand the performance of the system, we conduct ablation studies for our intra-turn and inter-turn relation prediction systems. We report the results of the **RST features**, both individually and in combination with the **argument relation features**. We also demonstrate the results of ablating this component on **our intra-turn system**, using only the **IMHO+Context Fine-Tuned BERT** model. Table 4.10 shows the comparison of the ablated components for intra-turn relation prediction. Our results demonstrate that the RST features are

Method	Precision		Recall		F-Score	
	Gold	Pred	Gold	Pred	Gold	Pred
Argument Relation Features	5.9	4.8	82.0	80.0	11.0	9.0
Argument Relation Features + RST Features	6.4	4.9	83.0	80.0	11.8	9.3
RST Features	5.1	3.8	80.0	77.0	9.6	7.2
QR+Context Fine-Tuned BERT	11.0	10.0	75.3	72.5	19.1	17.6
+ RST Features	11.0	12.2	79.0	75.5	21.2	19.1
+ Extractive Summarizer	16.0	14.5	**79.4**	**75.6**	26.8	24.3
Our Inter-Turn Relation System	**18.9**	**17.5**	79.0	74.0	**30.5**	**28.3**

Table 4.11: Ablation Experiments for Inter-Turn Relation Prediction with Gold and Predicted Claims

complementary to existing handcrafted features. Furthermore, we show that even intermediate fine-tuning BERT on the IMHO+context dataset is not sufficient and the RST discourse features are complementary.

Table 4.11 shows the comparison of the ablated components for inter-turn relation prediction. In this case, we demonstrate the effect of only using the **QR context fine-tuned BERT** model, as well as combining this model with **RST features** and the **extractive summarizer**. As with intra-turn relations, we find that RST features are complementary to handcrafted ones and are even complementary to the QR fine-tuning of BERT, which is not sufficient by itself. Our methods for candidate target selection obtain further improvement. The extractive summarizer component improves performance by approximately 5 points in F-score over only using the QR model with RST features (in both the gold and predicted cases) by reducing the search space of relation pairs. The full inter-turn relation system, which includes the constraint that targets may only be a source when targeting a main claim, we obtain another 4 point gain.

We might expect the RST features to perform better than what we observed. However, there are a few possible reasons for the under-performance. First, the RST parser was not trained on social media text and the performance of the parser is likely to decrease significantly due to the noise in this domain. Second, we used the RST parser in a non-traditional setting. Rather than identify the global structure of an entire post, we only considered pairs of propositions so that we could use a similar approach for both inter-turn relations (which would not have discourse structure) and intra-turn relations. Third, our model is limited by using only the relation at the root node of the RST parse, rather than any of the structure. A model that incorporates the full tree structure may obtain better performance.

Window Clipping We also conduct experiments showing the performance for intra-turn relation prediction when constraining the relations to be within a certain distance in terms of the number of sentences apart. Often, in persuasive essays or within a single post the authors use premises to back/justify claims they immediately made. As shown in Figure 4.17, this behavior is also reflected in our dataset where the distance between the two arguments in the majority of the relations is +1 (i.e. the premise immediately follows the claim).

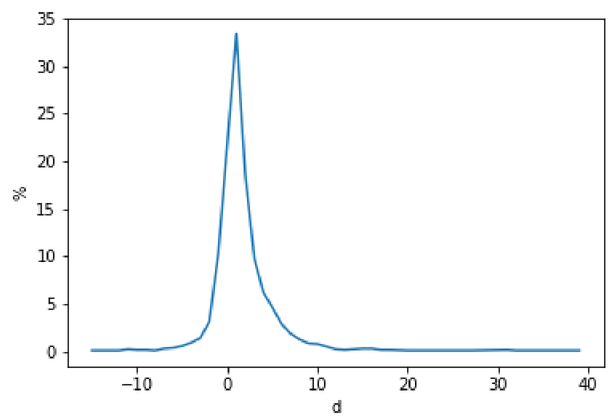


Figure 4.17: Distances *d* between Intra-Turn Relations

We thus limit the model's prediction of a relation to be within a certain window and predict "no relation" for any pairs outside of that window. Table 4.12 shows that this window clipping on top of our best model improves F-score by only limiting the context where we make predictions. As our models are largely trained on discourse context and the next sentence usually has a discourse relation, we obtain improved performance as we narrow the window size. While we see a drop in recall, the precision improves compared to our previous results in Table 4.8. It is also important to note that window clipping is only beneficial once we have a high recall, low precision scenario because when we predict everything at a distance of +1 as a relation we obtain low F-scores.

Method	Window	Precision		Recall		F-Score	
		Gold	Pred	Gold	Pred	Gold	Pred
All relations	0 TO +1	5.0	4.0	31.0	25.0	8.7	6.9
Best Model	0 TO +5	19.5	17.1	**70.0**	**67.0**	30.5	27.2
	0 TO +4	21.4	19.5	67.0	65.0	32.2	30.0
	0 TO +3	25.2	23.3	61.1	58.0	35.6	33.2
	0 TO +2	32.5	29.8	50.0	48.0	39.3	36.8
	0 TO +1	**41.5**	**39.1**	47.0	42.0	**44.1**	**40.3**

Table 4.12: Intra-Turn Relation Prediction with Varying Window Settings

4.2.4 Analysis

Finally, we conduct a qualitative analysis to examine the impact of our different system components, to study the impact of using the IMHO/IMHO+context and QR datasets as well as the role of RST relations.

Claim Detection To understand how using the IMHO dataset improved over the CNN and Task-Specific Fine-Tuning settings, we show examples that were incorrectly classified by the two baseline models but correctly classified by the IMHO Fine-Tuning. We retrieve the most similar example in the IMHO dataset to these misclassified samples according to TF-IDF over unigrams and bigrams. Table 4.13 presents the examples labeled by their dataset and the corresponding IMHO example. We find that the IMHO dataset contains n-grams indicative of claims, e.g. *"can be very rewarding," "should be taken off the market,"* and *"should intervene,"* demonstrating that the IMHO LM Fine-Tuning learns representations of claims based on discriminatory phrases. In fact, the CMV example is almost an exact paraphrase of the IMHO example, differing only in the phrase *"anecdotal evidence"* compared to *"my anecdotal experience."* At the same time, we find that many of the topics in these datasets occur in the IMHO dataset as well, such as *"public schooling"* and *"licence fees,"* suggesting that the language model learns a bias towards topics as well.

While empirical results indicate that IMHO Fine-Tuning helps in claim detection, we also investigated whether the language model introduces any bias towards types of claims. To this

Dataset	Sentence
WD	I send my daughter to public school but if I could afford to I would definitely send her to a nearby private school and not have to deal with lots of the problems in public schools.
IMHO	There is no telling that a private school will be better than public, that 's a parents choice, I pulled my kid from private school and went to public school that choice was made because the school we had access to was new and he excellent ratings and it was superior to the private school.
MT	That's why they should be taken off the market, unless they're unbreakable .
IMHO	Should be taken off the market.
MT	The Tv/Radio licence fee can only be required of all citizens/households equally.
IMHO	Radio 4 and Radio 6 music are pretty much worth the licence fee.
MT	Since, however, in Russia besides gas and oil only propaganda and corruption rule, the EU should intervene right away.
IMHO	Neither Russia or the EU should intervene in this case
CMV	Other than anecdotal evidence, I haven't seen anything to support this claim.
IMHO	I have personally seen no evidence to support this claim, but that's just my anecdotal experience .
PE	However, flourishing tourism in a place can be very rewarding in terms of local economy.
IMHO	It can be very rewarding.

Table 4.13: Sentences from each dataset and their nearest neighbor in the IMHO dataset

end, we also evaluated examples classified incorrectly by the model. Table 4.14 shows sentences that are predicted to be opinionated claims by our model but are actually non-claims. We note that a portion of these misclassified examples were premises used to back a claim which could be classified correctly given additional context. For instance, the second example from the MT data set in the table backs the claim *"It would be fair to make them into an Olympic event"* while the first example from the PE data set backs the claim *"There is no reason that governments should hesitate to invest in public transportation, a healthy, safe and economical way of transporting."* While discriminatory phrases like *"should"* or *"must be"* and comparative statements like *"much safer than"* or *"more ... than any"* are often indicative of claims, the lack of context may lead to incorrect classifications. Language modeling with additional context sentences or jointly modeling context (e.g. by predicting relations between claims and premises) may address these errors.

Dataset	Sentence
MT	If there must be rent increases , there should also be a cap to avoid nasty surprises
MT	Video games namely FIFA in my case , can fascinate young people for hours more intensively and emotionally than any sport in the world !
PE	Last but not the least , using public transportation is much safer than using private transportation
PE	In a positive point of view , when people without jobs have hand phones that have access to the internet , they will be able to browse the net for more job opportunities
CMV	Cheating is evidence , that *something* must be wrong

Table 4.14: Sentences which are actually non-claims but predicted as claims by IMHO Fine-Tuning

Dataset	Pair
IMHO	IMHO, you should not **quantify** it as good or bad. **Tons of people** have monogamous relationships without issue.
CMV	[how would you even **quantify** that] [there are **many people** who want close relationships without romance]
QR	[It might be that egalitarians, anti-feminists, MRAs & redpillers, groups that I associate with opposing feminism - might be in fact very **distinct** & **different** groups, but I don't know that] [**I do see** all four of these as **distinct** groups].
CMV	[I may have a different stance on seeing no **difference** between companion animals and farm animals.] [**I do see distinction** between a pet and livestock]
QR	[Of course you **intend to kill** the person if you draw your **weapon**, if you can reasonably assume that they have a **weapon**] [**I don't think** some of them would **start killing**].
CMV	[So i thought, why would a police officer even use **firearms** if he/she doesn't **intend to kill**?] [**I don't think**, police are allowed to **start killing** someone with their gun if they don't intend to .]

Table 4.15: CMV and Context Examples

Role of Context We retrieve examples from the **IMHO+context** and **QR** data using TF-IDF similarity to pairs of argument components from our data that were predicted incorrectly by pre-trained BERT but correctly by the respective fine-tuned model. The first two rows in Table 4.15 show a relation between a claim and premise in the IMHO+Context and the CMV data respectively while the last four rows show a relation between a claim and premise in the QR data and the CMV data. The model learns discriminative discourse relations from the **IMHO+context** data and correctly identifies this pair. The last four rows show rebuttal from the QR and the CMV data respectively, where the model learns discriminative dialogic phrases (highlighted in bold).

Discourse	Argument 1	Argument 2
Evaluation	The only way your life lacks meaning is if you give it none to begin with	Life is ultimately meaningless and pointless.
Antithesis	Joseph was just a regular Jew without the same kind of holiness as the other two	Aren't Mary and Joseph, two holy people especially perfect virgin Mary, both Jews? Wasn't Jesus a Jew?

Table 4.16: Predicted Discourse Relations in CMV

Role of Discourse We also provide examples that are predicted incorrectly by BERT but correctly by our classifier trained with RST features. For the first example in Table 4.16 the RST parser predicts an *Evaluation* relation, which is an indicator of an argumentative relation according to our model. For the second example the RST parser predicts *Antithesis*, which is correlated with attack relations [Musi et al., 2018], and is predicted correctly by our model.

4.2.5 Conclusions, Limitations, and Future Work

We showed that fine-tuning on context-appropriate datasets can be beneficial– the IMHO+context dataset of self-labeled opinionated claims and the QR dataset for dialogue interaction. As our labeled data set is relatively small we demonstrated how to use transfer learning by leveraging discourse and dialogue context to predict intra-post and inter-post relations, respectively. We also showed that predictions that take advantage of RST discourse cues are complementary to BERT predictions. Finally, we demonstrated methods to improve precision by identifying candidate targets.

However, while we model context using the IMHO+context and QR datasets, our predictions are made in isolation. In other words, we predict whether an argumentative component is a claim or a premise without considering the surrounding context. Furthermore, although our models of intra-post and inter-post relations consider two components, we could take advantage of additional context. End-to-end models such as the work of Eger et al. [2017] or Morio and Fujita [2018b] could be adapted to work with our fine-tuning framework. As claims and premises are inherently contextual [Lawrence and Reed, 2019], jointly modeling components and relations should result in additional improvement. Furthermore, our candidate target selection method is part of a pipeline, but using a reinforcement learning approach such as that of Chen and Bansal [2018] would allow us to recover from errors introduced during this stage. The same is true of our RST discourse features. As the RST parser is not perfect, we want to investigate additional features or models based on these trees that allow us to better recover from errors.

Additionally, our end goal is an end-to-end argument mining system at the token level. Our work here assumes that we have gold argument spans or that a proposition is an entire sentence. Ideally, the end-to-end model would identify whether a token is part of an argument and predict relations directly between all tokens. Exploring graph convolutional networks [Morio and Fujita, 2019] at the token level is one possible way to model this interaction.

Finally, future work could experiment further with language model fine-tuning on other sources of data. While we developed models for different tasks for the annotated data from Section 4.1, the models here do not predict the semantic types of claims and premises. One possibility is to examine how other datasets for dialogue tasks [Rosenthal and McKeown, 2015, Yang et al., 2019, Wang et al., 2019] can be leveraged in a multi-task learning or transfer learning framework to improve performance on our data.

4.3 Computational Models of Persuasion

In Section 4.1, we examined the impact of argumentative components and their semantic types and relations in terms of inter-post and intra-post interaction, showing that, for example, agreement and semantic coherence were key. In Section 4.2, we introduced methods to *predict argumentative components and relations* between them, with the goal of using the identified argumentative structure towards detecting persuasive arguments. Here, we take an alternative approach to *predicting persuasion*, modeling intra-post and inter-post interaction using neural methods. Rather than using our models of predicted argumentative structure, which may be unreliable for downstream tasks, we instead model intra-post interaction using a recurrent neural network [Hochreiter and Schmidhuber, 1997] and model inter-post interaction using a memory network [Xiong et al., 2016].

Our analysis and modeling of argumentative structure allows us to use aspects of claims and premises to determine when an argument will be persuasive, even without data labeled for persuasiveness to use for training a model. However, we also need to consider the converse: if we have data labeled for persuasion, we may be able to infer aspects of structure or semantics from a model trained to predict persuasion. While our models of argumentative structure have been shown to be effective, we may not always have labeled data for a particular domain. Furthermore, additional rhetorical aspects of persuasion may not be captured by our coarse-grained labels. We thus desire an approach to predicting persuasion that requires only text and additional automatically-derived features.

Predicting persuasion is a difficult task as in addition to modeling aspects of discourse and dialogue, it requires modeling world knowledge and reasoning. Evaluating the use of world knowledge or logical reasoning in a natural language argument is beyond the scope of our current capabilities, so we model the interaction between the original post and response and the sequence of arguments used in the response. In terms of intra-post interaction, we model these relations at a *micro level* using a recurrent neural network over the sentences in the original post and response, representing the sentences using word embeddings as well as discourse relations, semantic frames, and structural features such as paragraph breaks. As a result, the recurrent nature of this method allows us to capture the coherence of the argument and the additional features allow us to capture and evaluate the use of rhetorical moves such as agreement and the judicious use of discourse connectives. Given the sentence representations obtained from the neural network for both the original post and response, we then model inter-post interaction at a *macro level*, modeling the interaction between the entire original post and each sentence in the response using a memory network.

In Section 4.3.1, we discuss how we model this task and use data from Change My View for our experiments. We then describe how we **model intra-post and inter-post interaction** using a recurrent neural network and a memory network, respectively (Section 4.3.2). We then provide empirical evidence that the change in beliefs of online posters is not just caused by novel words and concepts but by their presentation in terms of ordering and interaction, showing in Section 4.3.3 that our models are effective compared to previous work and other baselines. Finally, we provide an analysis and ablation studies in Section 4.2.4 to better understand why our model of sequencing and interaction is effective, showing that we even outperform novice human annotators and illustrating the difficulty of this task.

4.3.1 Data

As with our annotation of arguments in Section 4.1.2, we use the Change My View subreddit described in 3.1.1. In previous work, Tan et al. [2016] collected threads (full discussion trees) submitted between 2013/01/01 and 2015/09/01, and segmented this data into submissions before and after 2015/05/08. This process resulted in 18,363 and 2,263 discussion trees, respectively, for train and test. As the data is self-labeled by posters for persuasion using the Δ character, no human annotators are required.

We consider three tasks. The first is our primary task of interest – **influence** prediction where, given a post and response, we attempt to predict whether the user changed their view. This task is most similar to the scenario where we have a number of retrieved or generated candidate arguments and we would like to evaluate them and select the most convincing. For this task, we extract dialogs from the discussion trees crawled by Tan et al. [2016]. We take a similar approach to our annotation of arguments in Section 4.1.1– we select dialogues by following paths in the discussion tree where only the original poster and one responder interacted. We automatically label positive and negative examples based on whether the path terminates with and without the original poster providing a Δ, respectively. We extract only one path per response to the original poster by following the left-most path in a depth-first search and allowing a single unique response per path. Each datapoint is then an original post and attempted persuasive response, where responses are one or more sequential posts from the same commenter. For training, we require every original post in the data to have at least 1 positive and 1 negative response. The resulting training set has 19516 examples (14849 negative and 4667 positive). The test set contains 2465 examples (1836 negative and 629 positive).

The second and third tasks are the same as previous work [Tan et al., 2016]. For the **pairwise** task, we predict which of two responses to the same original poster changed their view, where the two responses are controlled for topic by Jaccard similarity. This arrangement allows us to study persuasive aspects of an argument while controlling for content. The third task is **malleability** prediction, where the goal is to predict persuasion given only the original post and no responses. The purpose of this task is to evaluate when an argument is worth responding to – the original poster may be resolute and unable to argue in good faith.

Tan et al. [2016] distinguished two cases of the path-based prediction: predicting a delta from only the initial post in the response (termed the *root reply*) and including all posts in the response (termed the *full path*). For our experiments, at minimum the root reply and/or original post are available. Including the full path allows us to study structural features of the dialogue, whereas the root reply by itself may provide other indicators of convincingness (e.g. their style or coherence).

4.3.2 Methods

To build models for each of the tasks described in Section 4.3.1, we use a deep learning approach where we represent words and sentences hierarchically to obtain a post-level representation of the entire counter-argument and combine this representation with global features used in previous work [Tan et al., 2016] to make a prediction. First, we obtain a **sentence representation** using a weighted average of embeddings for words, frames, and discourse relations combined with structural features such as post breaks (see Figure 4.18). Next, we model the sequencing of sentences using an LSTM [Hochreiter and Schmidhuber, 1997]. We obtain these contextual sentence-level representations for both the original post and response. One possibility is to learn the interaction between the original post and response using an attention mechanism for each pair of sentences [Chen et al., 2017b]. Instead, we use a memory network to learn this interaction and condition each response sentence on the entire original post (see Figure 4.19), as this model has demonstrated its effectiveness at modeling context on related tasks such as opinion recommendation [Wang and Zhang, 2017a]. In this way, we obtain a **document representation** by combining the weighted

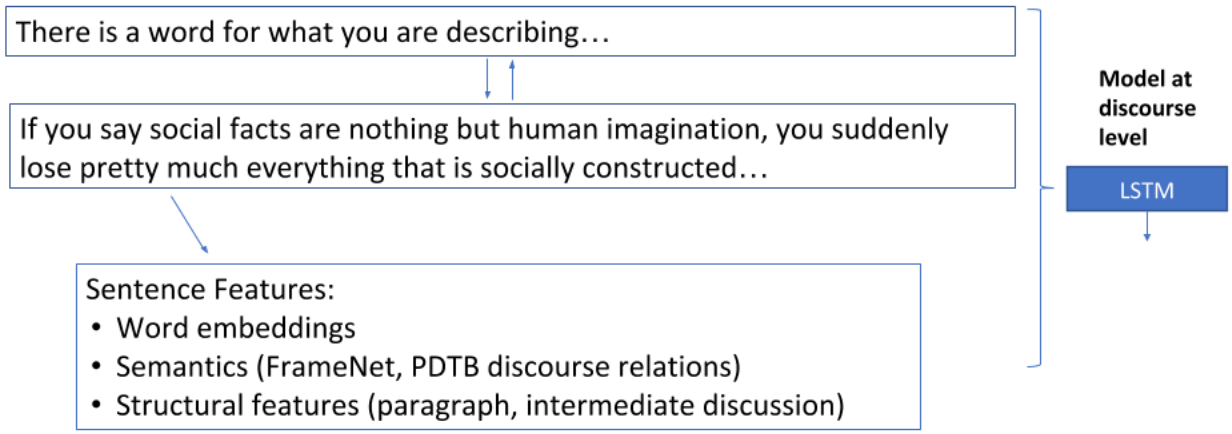


Figure 4.18: Sentence Representation for Persuasive Influence Network. Each sentence is first encoded using word embeddings and semantic and structural features. Then a contextual representation of the sentence is obtained by an LSTM.

average of the response sentences with global features based on the overlap of content and stop words between the original post and response [Tan et al., 2016].

Sentence Representation We create a sentence representation $\mathbf{r}_s$ by combining features from words, semantic frames, and discourse relations. We first represent each sentence by a weighted average of its word embeddings. Given a sentence at index s with T words and word embedding $\mathbf{x}_{s,t}^{word}$ for $t \in [1, T]$, the vector for s is:

$$\mathbf{v}_s^{word} = \sum_{t \in [1,T]} \alpha_{s,t}^{word} \mathbf{x}_{s,t}^{word} \tag{4.1}$$

Similarly, we add embeddings for semantic frames. Recent research has shown that neural models of documents have difficulty learning discourse and dialogue structure without explicit supervision [Reed et al., 2018]; we thus hypothesize that word embeddings alone are not sufficient to capture this aspect of intra-post and inter-post interaction. The FrameNet [Ruppenhofer et al., 2006] model of frame semantics provides a structure for events, relations, and objects and how they interact. It also provides a way to model social interactions that are not captured by discourse structure or explicitly expressed in words such as agreement and disagreement. For example, the verb "agree" may take the "Compatibility" frame, which is shared with similar verbs. Our analysis from Section 4.1.4 showed that agreement is effective in convincing arguments and ideally we would incorporate this into models of persuasive influence prediction. We use a FrameNet parser [Das et al., 2010] to predict the labels for lexical units and represent frames as the weighted average of the labels:

$$\mathbf{v}_s^{frame} = \sum_{l \in [1,L]} \alpha_{s,l}^{frame} \mathbf{x}_{s,l}^{frame} \tag{4.2}$$

where $\mathbf{x}_{s,l}^{frame}$ is the embedding for the lth frame and L is the total number of frames. We create an embedding for each lexical unit that evokes a specific frame, by randomly initializing a vector for each observed class in the dataset and updating during training.

Each attention weight α_{sj}^k is calculated for each $\mathbf{x}_{sj}^k$ for $k \in \{word, frame\}$ and $J \in \{T, L\}$, respectively, where T is the number of words and L is the number of frames:

$$\alpha_{sj}^k = \frac{\exp({\mathbf{u}_{s,j}^k}^T \mathbf{q}_k)}{\sum_{a \in [1,J]} \exp({\mathbf{u}_{s,a}^k}^T \mathbf{q}_k)} \tag{4.3}$$

and $\mathbf{u}_{s,j}^k = \tanh(W_k \mathbf{x}_{s,j}^k + \mathbf{b}_k)$ and $\mathbf{q}_k$ is a parameter vector.

Finally, we augment the sentence representation by incorporating embeddings for discourse structure, again hypothesizing that the document-level neural model is not sufficient. Previous work [Tan et al., 2016] used patterns of connectives such as "but-however-because" as features, but noted that these models suffered from low recall. Thus, modeling implicit discourse relations should improve coverage as implicit discourse is not explicitly captured by the remainder of the model. Here we use the Penn Discourse Tree Bank (PDTB) model of discourse relations, which defines a shallow local view of semantic relations between adjacent segments. We use the end-to-end model of Biran and McKeown [2015] to tag PDTB relations rather than alternatives such as RST so that we can incorporate shallow structure into our LSTM. We represent the second-level discourse classes (e.g. Contingency/Causal and Comparison/Concession) for each inter-sentence relation as an embedding for sentence s as $\mathbf{v}_s^{inter}$, indicating the relationship between s and $s-1$.

The final sentence representation is then determined by concatenating each component of the sentence :

$$\mathbf{v}_s = \left[\mathbf{v}_s^{word}; \mathbf{v}_s^{frame}; \mathbf{v}_s^{inter}\right] \tag{4.4}$$

In order to model features such as post structure, we insert a single "intermediate discussion" token in between posts so that the LSTM can learn to identify the start and end of a post.

Given $\mathbf{v}_s$, we could use this representation of each sentence for the input at each timestep of an LSTM, or model feature interaction by applying a multi-layer perceptron (MLP) to $\mathbf{v}_s$. Instead, we follow previous work in hierarchical language modeling [Kim et al., 2016] and allow the model to decide whether to *carry* features directly to the next layer, in order to allow for interaction between the word, discourse, and frame semantic features derived during this step. We thus obtain our sentence representation by feeding $\mathbf{v}_s$ into a highway network [Srivastava et al., 2015]:

$$\mathbf{r}_s = \mathbf{t}_s \odot \mathbf{z}_s + (1 - \mathbf{t}_s) \odot \mathbf{v}_s \tag{4.5}$$

where $\mathbf{z}_s = g(W_h \mathbf{v}_s + \mathbf{b}_h)$, a hidden representation of the original vector with a non-linearity g, and $\mathbf{t}_s = \sigma(W_t \mathbf{v}_s + \mathbf{b}_t)$, a prediction of whether to use the original features. The highway network is a mixture of the hidden representation of the vector given by the MLP and the original vector, where the model learns the weight vector $\mathbf{t}_s$. Thus, because of the learned weight $\mathbf{t}_s$, the model decides how to interpolate between the hidden representation and the original vector.

Document Representation Given the sentence representation $\mathbf{r_s}$, where s is the index of the sentence, we model the document as a bi-directional LSTM with an attention mechanism over the sentences. We first obtain the LSTM states for all sentences $\mathbf{h}_s^r$ in the response and $\mathbf{h}_s^{op}$ in the original post.

One possibility is to use a bi-directional LSTM for $\mathbf{h}_s$ over the sentences from the reply only. However, this would only allow the attention mechanism to consider the response, rather than the

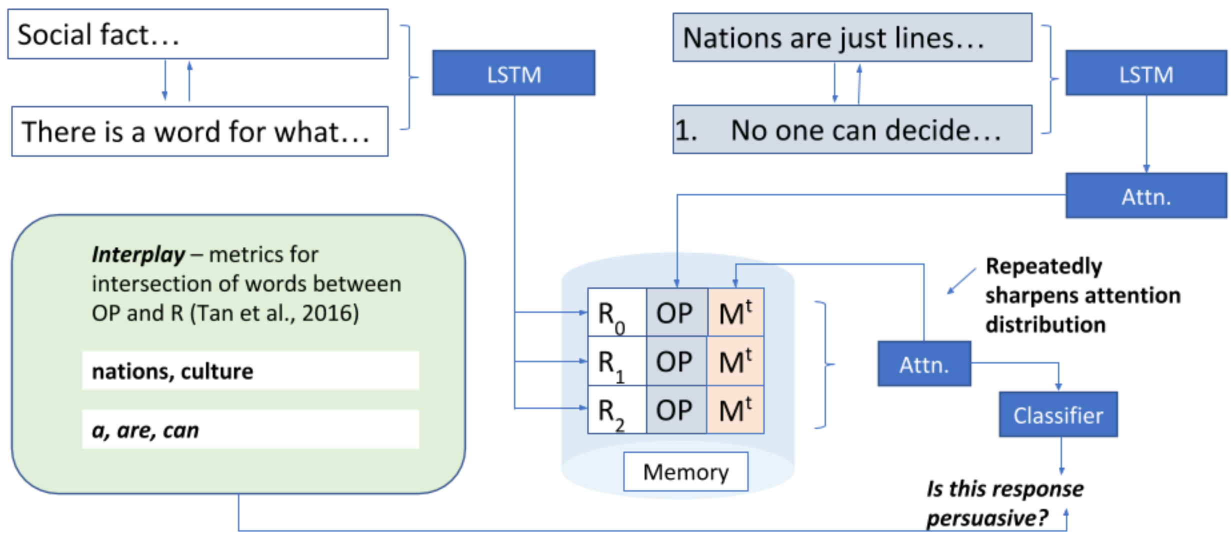


Figure 4.19: Persuasive Influence Network. The contextual representations for all original post (OP) and response (R) sentences are obtained from an LSTM using features from Figure 4.18. The OP post representation and R sentence representation are then combined and the attention over the R sentences is repeatedly sharpened using a memory network. The final representation of the R post is then the attention-weighted combination of all the R sentence representations, which is combined with global features (e.g. interplay) to make a prediction.

context of the original post. We thus include information about the original post using a dynamic memory network, which has been effective in modeling context [Xiong et al., 2016, Wang and Zhang, 2017a], to iteratively find abstract representations using information from both the original post and the response. While we could model the interaction between each sentence of the post and response using an alternative attention mechanism [Chen et al., 2017b], we hypothesize that the use of the memory network allows us to better model the hierarchical structure we observed in Sections 4.1 and 4.2. In this framework, the model attends over the sentences in the original post, learning to identify the most salient sentences. This attention mechanism can be considered a "soft" representation of the extractive summarization approach we took in Section 4.2.2 to identify candidate targets. Then, the memory network allows us to identify likely source-target pairs (to use the terminology from Section 4.2.2). One advantage of the memory network is that it uses multiple "hops" to learn the attention over the response sentences, iteratively re-computing the attention mechanism to find "sharper" peaks. As our experiments in Section 4.2.3 showed, only a few pairs have relations; the memory network thus allows us to increase the probability assigned to a small number of source-target pairs while decreasing the probability assigned to others.

Given the LSTM state $\mathbf{h}_s^{op}$ at sentence s in the original post, we first create a representation $\mathbf{h}^{op}$ for the entire original post by attending over each hidden state, similar to previous work [Yang et al., 2016]:

$$\mathbf{h}^{op} = \sum_{s \in [1,S]} \alpha_s \mathbf{h}_s^{op} \tag{4.6}$$

where S is the number of sentences in the document and attention is calculated by applying an MLP to the hidden state, $\mathbf{u}_s = \tanh(W_s \mathbf{h}_s^{op} + \mathbf{b}_s)$, before calculating the probability distribution over sentences (using $\mathbf{q}^{op}$ as a learned parameter vector):

$$\alpha_s = \frac{\exp(\mathbf{u}_s^T \mathbf{q}^{op})}{\sum_{i \in [1,S]} \exp(\mathbf{u}_i^T \mathbf{q}^{op})} \tag{4.7}$$

$\mathbf{h}_s^{op}$ is concatenated with $\mathbf{h}_s^r$ and a "memory representation" $\mathbf{v}^t$ to create the input representation: $\mathbf{h}_s^t = [\mathbf{h}_s^r; \mathbf{h}^{op}; \mathbf{v}^t]$. By allowing the attention mechanism to consider the context and the entire response, the model is able to more accurately predict which sentences are important. Initially, the memory $\mathbf{v}^0$ at $t = 0$ is set to the average of the hidden states: $\sum_{s \in [1,S]} \mathbf{h}_s^r / S$. Then, after up to t "hops," the memory $\mathbf{v}^t$ is set to $\mathbf{h}^{t-1}$. Similar to Equation 4.6, we obtain an attended document representation of the response:

$$\mathbf{h}^t = \sum_{s \in [1,S]} \alpha_s \mathbf{h}_s^t \tag{4.8}$$

again using a learned parameter vector $\mathbf{q}^r$ to compute the attention:

$$\alpha_s = \frac{\exp(\mathbf{u}_s^{rT} \mathbf{q}^r)}{\sum_{i \in [1,S]} \exp(\mathbf{u}_i^{rT} \mathbf{q}^r)} \tag{4.9}$$

We could use $\mathbf{h}^0$ as the final document representation $\mathbf{h}$, but in practice multiple iterations have been more effective [Xiong et al., 2016], which we validate empirically by experimenting with different values of t to determine $\mathbf{h} = \mathbf{h}^t$.

Finally, this document representation $\mathbf{h}$ is then passed through an MLP to make a binary prediction of influence, which is combined with global features ϕ (where β is a learned parameter vector) derived from the interaction between the original post and response:

$$y = \sigma\left(MLP(\mathbf{h}) + \beta^T \phi\right) \tag{4.10}$$

In our experiments, the global features we used are the *interplay* features of Tan et al. [2016]. In their experiments, their best-performing features were derived from the overlap between the original post and the response. They derived 12 features from 4 similarity scores (common words, similar fraction in reply, similar fraction in OP, and Jaccard score) and 3 subsets (all words, stop words, and content words). Tan et al. [2016] found that high overlap in stop words and low overlap in content words is likely to make an argument more convincing. This may be because good arguments use “entrainment” [Brennan, 1996] and thus the same style and also provide novel content rather then repeating the same content words back to the original poster. We hypothesize that these features are complementary to those of our model, which captures coherence, social interaction, and post structure.

4.3.3 Experiments and Results

We train and evaluate on the data described in Section 4.3.1 for each of the persuasive influence, pairwise (balanced), and malleability tasks and with the root reply or full path response where applicable.

The results of our experiments on the held-out test set are shown in Tables 4.17, 4.18, and 4.19 for each of the influence, pairwise, and malleability prediction tasks, respectively. For the pairwise and influence prediction subtasks, we report results for both the root reply and full path options and we compare models using sentences from just the response (R) and the response plus the original post (R+OP).

We present results using only words (word-LSTM) and words, frames, and discourse relations (all-LSTM). For the pairwise and influence tasks, these models consider the response only. For the malleability task, these models consider the original post. When the response and original post are both provided, we use the memory network described in 4.8 (all-LSTM+memory) to capture the additional context. We also provide results for interplay (IP) features combined with our model (all-LSTM+memory+IP), with the features as ϕ in Equation 4.10.

Baselines We provide baseline models from previous work [Tan et al., 2016], trained using logistic regression on features from just the response (*bag-of-words*) and from the response plus the original post (*interplay*). The interplay (IP) provides a strong baseline because we might expect there to be significant overlap between the posts if users are imitating the writing style of the original poster in order to be more persuasive. In addition, we provide a *bag-of-words* (BoW) baseline. We remove words occurring less than 5 times and L2-normalize term frequency vectors. We also compare our model to a strong baseline – an LSTM over averaged word embeddings with attention (word-LSTM).

Hyper-parameters and Optimization We use binary cross-entropy as the loss function and stochastic gradient descent with a mini-batch size of 100 and Nesterov momentum with a coefficient of 0.9. Word embeddings are initialized with pre-trained 300-dimensional GloVe vectors [Pennington et al., 2014]. Out-of-vocabulary words are randomly initialized and optimized during training. We stop training after 30 epochs and perform early stopping on a validation set (setting aside 10% of the training sets described in Section 4.3.1). The document weights β in Equation 4.10 were pre-trained using a logistic regression classifier.

We experimented with different settings for various hyper-parameters. For the recurrent and hidden dimensions, we tested values of 50, 100, 200, and 300. For dropout [Srivastava et al., 2014] and word dropout [Iyyer et al., 2015], we used values of 0.25, 0.5, and 0.75 and determined whether to use 1 or 2 hidden layers. We use ReLU as the non-linearity in Equations 4.5 and 4.10, according to our experiments on the validation set. We evaluated the number of iterations for the memory networks and found that performance increases up to 3 iterations and begins decreasing after 3. We limit the maximum length of each sentence to 32 words and the maximum length of a post to 40 sentences. Words occurring fewer than 5 times in the training set (including the original post, title, and response) were removed.

	Model	***Root Reply*** ***Acc.***	***AUC***	***True F-score***	***Full Path*** ***Acc.***	***AUC***	***True F-score***
R	BoW	60.4	68.9	47.1	61.9	72.8	50.3
	word-LSTM	71.2	70.5	48.7	72.9	75.1	52.7
	all-LSTM	**72.5**	**70.8**	**48.9**	**75.1**	**75.5**	**53.0**
R+OP	IP	70.5	74.8	52.1	72.7	76.7	54.6
	all-LSTM+memory	75.0	74.9	53.1	74.3	77.3	55.4
	all-LSTM+memory+IP	**77.2**	**79.5**	**58.0**	**81.0**	**82.1**	**60.7**

Table 4.17: Results of Influence Prediction Task

Discussion Our results in Table 4.17 show that the LSTM models significantly outperform all baselines, especially when combined with the interplay features. In the **influence** prediction task, the best model using only the response (all-LSTM) outperforms the BoW baseline in both the root reply and full path cases ($p < 0.001$ by a randomized permutation test). Given the response and the original post, the best model (all-LSTM+memory+IP) outperforms the IP baseline in both cases ($p < 0.001$). The difference between the best model and the baseline is also larger in the full path case when compared to the root reply case. This is not surprising, as many responses in our dataset contain only a single sentence, often a clarifying question, so the model is unable to benefit from sequential information when only the root reply is included. We also observe that modeling the context of the original post helps in both scenarios, but the context is more important in the root reply case, obtaining around a 4 point increase from all-LSTM to all-LSTM+memory compared to 2-3 points in the full path case. As the model has limited content to work with in the root reply case it is most likely taking advantage of features in the original post. Additionally, it is surprising that interplay is such a strong baseline, especially in the root reply case. Our all-LSTM+memory model does not significantly outperform the interplay features alone but does provide a complementary approach.

For the **pairwise** prediction task, we obtain better performance on accuracy ($p < 0.001$ by McNemar's test, comparing all-LSTM to BoW in both the root reply and full path cases, and

$p < 0.01$ comparing all-LSTM+memory+IP to the IP baseline). By controlling for topic in the pairwise dataset, individual words have less influence. Even though the model contains shallow structural features, word embeddings are a central part of the model, so the fact that the model performs well on pairwise prediction even with controlling for topic similarity suggests that the ordering of the document is key. Furthermore, we do not see significant improvement by including context in the pairwise task, which may indicate that the model is learning a bias for features of the original post rather than interacting with the response.

Finally, we would expect BoW to do well on **malleability**, as Tan et al. [2016] showed that common words associated with openness or stubbornness were strong features. However, we see significant gains from sequential models ($p < 0.05$ by a randomized permutation test for all-LSTM), suggesting the ordering of arguments provides some indicator of how and whether they can be convinced.

	Model	*Root Reply*	*Full Path*
R	BoW	59.6	62.3
	word-LSTM	67.0	70.8
	all-LSTM	**67.5**	**71.5**
R+ OP	IP	65.2	69.2
	all-LSTM+memory	67.7	71.6
	all-LSTM+memory+IP	**69.0**	**71.9**

Table 4.18: Accuracy for Pairwise Prediction Task

Model	*Acc.*	*AUC*	*True F-score*
BoW	51.6	53.3	48.1
word-LSTM	57.7	55.5	**56.5**
all-LSTM	**58.4**	**57.2**	53.2

Table 4.19: Results of Malleability Prediction Task

Model Component Ablation We present additional results in Table 4.20 on the full path task for influence, with certain model components from the all-LSTM model ablated to assess their contribution to modeling the sequence of reasoning. We remove the highway network component of the model, indicated in the table as *no highway*, and instead directly use the concatenated embeddings

Model	*Accuracy*	*AUC*	*True F-score*
all-LSTM	**75.1**	**75.5**	**53.0**
no highway	70.1	74.9	52.6
no lstm	68.8	73.2	50.3
no attention	66.6	74.5	51.3
discourse only	54.6	63.6	43.5
frames only	43.3	66.4	44.2

Table 4.20: Model Component Ablation

$\mathbf{v}_s$ as the input to the bi-directional LSTM. We also remove the bi-directional LSTM from the model, indicated in the table as *no lstm*, and instead take a weighted average of all the embeddings $\mathbf{v}_s$. Finally, we remove the attention mechanism over the LSTM states (*no attention*) and instead average the LSTM states over each timestep. We also present the impact of discourse and frame embeddings when included in the model without the other embeddings.

As demonstrated in Table 4.20, the sequential nature of the LSTM contributes to the overall performance of the model. Compared to the full model, the model without an LSTM (which considers the ordering of the content provided) does 2-3 points worse in AUC and F-score, showing that modeling the sequence of arguments helps in predicting persuasion ($p < 0.01$ by a randomized permutation test). We also obtain improvement by including the highway network and the attention mechanism ($p < 0.05$). Removing the highway or the attention component costs the model 0.5 to 1 point of performance. Without the highway layer, the neural network can only consider the sentence features individually and not the interaction between components. Without the attention layer, the model is unable to determine which parts of the sequence are most important to weight in the final prediction. Finally, the frame and discourse embeddings perform poorly on their own, but contribute to the overall model.

In summary, our finding in Section 4.1.3 that agreement and coherence are more effective in winning arguments is supported by our ablation studies. The features based on semantic frames are complementary to those based on word embeddings alone, as shown when we ablate the highway network. Furthermore, the components of the model that capture coherence – the LSTM and the post structure – are not present when using only the interplay features, and we observe a significant performance drop when the LSTM is ablated.

4.3.4 Analysis

We present additional analysis of our model to better understand the features of persuasive influence as well as the areas in need of improvement. An additional goal of this analysis is to determine whether we can provide additional support for our observations in Section 4.1.3, even though we have no labeled data for argumentative components and semantic types.

Model Evaluation We conduct an analysis of the model by examining the attention weights, first at the sentence level, then at the document level.

One advantage of this model is that we can easily see which words, frames, or discourse relations are prominent features according to the attention-based weighting. For the influence task, highly-weighted words include terms such as *objectified*, *stereotyped*, *thesaurus*, and *linguist* which may just indicate that people have strong opinions on these topics. Highly-weighted frames, however, include *research* and *medical_professionals*, which may indicate users providing evidence, or *confronting_problem* and *suasion* (attempts to persuade), which may indicate social interaction. In the malleability case, highly-weighted words include *greetings* and *brigading* (a Reddit term for a group of users coordinating to downvote certain posts), which indicate social aspects of persuasion. Other highly-weighted words include terms such as *protectionism* and *anarcho* (a word in the context of anarcho-capitalism), which is unsurprising as politics is a controversial topic. Highly-weighted frames include social cues such as *contrition* or *hostile_encounter*, which may indicate susceptibility or resistance to persuasion, respectively.

We also conduct a qualitative analysis to evaluate the impact of the *sentence*-level attention weights. We present results showing human judgments of the most important sentences in the response and we compare the results of this annotation task to the attention weights output by the model, as in the work of Ghosh et al. [2017]. We designed an Amazon Mechanical Turk (AMT) task to conduct our experiments. We provide the annotator with an original post and the sentences in the reply, so that the annotators have access to the same data as the model for a fair comparison. The annotator is asked to indicate the "most important" sentences in the response. They are then required to select at least one sentence but may select the entire response. We use a subset of 80 test examples for our experiment and limit the length of the original post and reply to be between 3 and 10 sentences to simplify the task for the annotators. This results in 36 positive and 44 negative examples. Each HIT contains one task and 5 annotators were required for each task. Only Master-level annotators were selected and workers were paid 10 cents per HIT, in accordance with the U.S. federal minimum wage.

We first compare the sentence-level weights of the all-LSTM+memory model to the annotators' selections. We find that 32% of the time the highest-weighted sentence from the model is the sentence where the most annotators agree that the sentence is important. We also find that 35% of the time, the highest-weighted sentence from the model is the second-most important sentence from the annotators. Of the remaining 33%, the model selects the first sentence 60% of the time, indicating a bias towards the beginning of the text. Overall, a baseline method of always weighting the first sentence the highest would achieve 20% accuracy compared to the annotators. In this subset of data, the average length of the positive posts is 6.25 sentences and the average length of the negative posts is 6.27 sentences. Even though the posts are the same average length, we find that for positive responses, the Turkers selected 19% of all sentences whereas for negative responses, they selected 16%, indicating that positive responses contain more important content.

We also provide an example of attention weights along with the predictions made by annotators in Table 4.21. The title is *"College is not unaffordable in the US."* and the original post is also provided. The full text of a response that received a delta and one that did not are both shown as well, segmented into sentences. The "Labels" column indicates the percentage of annotators that voted for that sentence and the "Attn" column indicates the probability assigned to the sentence by the model. The Attn column will thus sum to 1 but the Labels column will not, so we compare the relative ranking of each sentence. The top-ranked sentence by the annotators is highlighted in bold. In both cases this sentence could act as a summary for the entire argument. However, the attention weights in this example do not reflect this ranking. The overall prediction for both responses was incorrect and a correct prediction may only be possible with world knowledge (about the value of money).

CMV: College is not unaffordable in the US
OP
The expense of a higher education has been a hot button issue for a number of years but the fact that students take on huge loans and graduate with huge sums of debt is a function of their own suboptimal decisions. According to google the average tuition for private colleges is $31,231 per year. Assuming the worst case and a student pays full sticker price, that's around 125k in debt. But with almost all private schools, there are plentiful scholarship and grant opportunities. Even for students who support themselves (meaning no financial help from their parents), these options make the 125k number far less. This forgets the fact that according to the same source, students can attend an in-state university for just shy of 10,000, for a far more affordable (and worst case) of 40k. And this is without considering alternatives to college.

Positive	***Attn***	***Label***
Are you arguing that collage is affordable, or more affordable than people imply?	0.28	0.2
Because while I would agree that there is likely some exaggeration, for many people it is completely unaffordable.	0.29	0.4
Not everyone gets the best case scenario, and if you make less than $30000 a year, then paying minimum of a third of a years pay on education is not feasible.	0.23	0.6
And I don't know how considering alternatives to collage is an argument for the affordability of collages; yes collage is cheep if I do not go to it and take an apprenticeship instead, but I don't know what it would have to do with this discussion.	0.2	0.4

Negative	***Attn***	***Label***
My family made "too much" for FAFSA aide but too little to afford me much assistance with college prices.	0.19	0.6
I went to a school where I was given a full academic scholarship, which included room and board.	0.18	0.4
In order to afford additional fees / books / transportation I still had to take out a Stafford loan every year.	0.16	0.4
On top of that, the government decided that the room and board part of my scholarship qualified as "income", and I then owed the IRS money come tax return time for each of my four years.	0.15	0.4
I'll still be paying off these loans for a few years.	0.16	0
My point: Even with the "best case scenario" of a full scholarship, college still poses a significant financial burden.	0.16	0.4

Table 4.21: Attention Weights and Human Annotations from a Positive and Negative Example

Human Performance We also carry out an evaluation of human judgments to compare performance. We set up an experiment on Figure Eight where we ask annotators to view discussions from Change My View. For each discussion thread, we display the original post and title, then display one positive argument and one negative argument in a random order. For each argument, we display all posts from the author of the root response so that, similar to our attention experiments, the annotators have access to the same data as the model. This is equivalent to the "full path" task in our experiments.

Model	***Pairwise***	***Influence***
Annotators	54.84	57.14
all-LSTM+IP	71.99	63.00

Table 4.22: Human Performance

First, for each argument, we ask the annotator whether they believe the original poster would find the argument convincing. Then we ask annotators to rank the arguments, to compare to the pairwise accuracy task. We instruct the annotators to read the original post and both arguments before answering any questions. For quality control, for each of the three questions we require each annotator to provide a justification of at least 20 words for their decision. Justifications that did not meet this requirement or were clearly spam had their judgments removed from the dataset. As an additional quality control measure, we require annotators to spend at least 300 seconds on each discussion; consequently they were paid 75 cents per discussion. Annotators are required to give three judgments per thread and we annotate a total of 200 discussion threads. Results are presented in Table 4.22, showing the majority vote of the annotators along with our model performance on the same subset of data.

It is not surprising that human annotators struggle with both the pairwise prediction task and the influence prediction task. If humans were better at predicting when a post would be persuasive, we would likely see more persuasion in our dataset. Our models significantly outperform human annotators on both tasks. One key distinction is that the annotators received no training in what makes a successful argument, whereas our models are trained on thousands of documents. An expert in persuasive writing may perform very well at this task so we can only claim that our

		Human		*Model*	
Category	***%***	***P***	***I***	***P***	***I***
Government	29	76.3	55.1	64.4	58.5
Sociology	23	71.7	53.3	80.4	68.5
Morality	11	72.7	63.6	77.3	68.2
Economics	9	50.0	50.0	72.2	58.3
Politics	8	62.5	56.3	68.8	62.5
Science	6	66.6	66.6	66.6	62.5
Culture	5.5	54.5	45.5	54.5	63.6

Table 4.23: Error Analysis on Categorized Data (P: Pairwise I: Influence %: Percentage of Data in Category)

model is better than novice annotators.

Error Analysis In order to discern the areas of improvement for models of persuasive influence, we categorize examples into several categories to see how our models and the human annotators fare. Then we report performance on each category. We divide all posts in the human-annotated subset into seven broad categories: government (what laws should be implemented), sociology (behavior of groups or discussion of social issues such as feminism), morality (judgments of right and wrong), economics (personal or group decisions to maximize utility), politics (what political parties and candidates should do), science (questions with objective, measurable answers such as whether vaccines are effective), and culture (books, music, games, etc.). Each post is categorized by the first author and any post not clearly belonging to a category is discarded.

In an example of the politics category, an original poster writes: *"There is no practical reason for any individual to vote in national elections. By 'practical reason,' I mean a reason that motivates you to vote by ascribing a cause-effect ... This is a classic example of a collective action problem."* In a winning argument, a user writes: *"Just because it's incredibly unlikely that your vote will make a difference doesn't mean it's never going to happen. ... Depending on a person's valuation of costs and potential benefits, this could very well be enough."* In contrast, another user writes an unconvincing argument: *"The same ballot for Presidential and Congressional elections will also have a number of other state and local positions and issues ... Then you are putting in* *a very low amount of effort for a very low amount of impact."* On this example, the human annotators correctly predict the positive response but not the negative one whereas our model correctly predicts both.

The overall results for accuracy are reported in Table 4.23. Overall our models perform best on topics in sociology and morality and have issues with discussions in government and economics. We observe that in CMV the former tend to be more emotional (for example, in response to the original poster writing *"Weinberg was wrong when he said that 'for good people to do evil things, that takes religion' "* another user writes *"I think that someone isn't a good person if they have an ideology I disagree with"*) while the latter tend to be more empirical (for the topic *"Countries should have a 'no confidence' vote in elections if they want to increase turnout, while achieving a better understanding of the public's perception of the political climate"*, another poster responds with facts: *"The US state of Nevada has had a choice called 'none of these candidates' since 1975."*) As the empirical arguments often require world knowledge we would expect our models to struggle in this area. Conversely, our models may pick up on sequential arguments alternating between emotion and logic in other categories. For example, *I think that someone isn't a good person if they have an ideology I disagree with* is followed by *"I think nationalists are bad, fascists are bad and so on."* The model correctly identifies the post with these arguments as not receiving a delta, which may be due to the sequence of simplistic, emotional language used. Finally, compared to human performance, our models are worse or at the same level in government and science, suggesting that world knowledge may again be the distinguishing factor.

Our analysis here is aligned with our analysis in Section 4.1.3, which showed that pathos alone was not indicative of a winning argument and consecutive arguments of the same semantic type (both pathos and logos) were more coherent and therefore more effective. Furthermore, the weak performance in categories such as government and economics that likely require world knowledge suggests an area for improvement. One possibility for incorporating world knowledge is by leveraging an external resource such as Wikipedia. We thus hypothesize that automated fact-checking methods or computational methods for identifying or retrieving supporting evidence would allow for the verification of claims and premises and that factual arguments are more likely to be effective.

4.3.5 Conclusions, Limitations, and Future Work

We showed that the ordering of and interaction between arguments is important for persuasion by presenting a neural model of persuasive influence, modeling words, PDTB relations, and FrameNet semantic frames. We demonstrated statistically significant improvements over previous work on predicting persuasion by using features representing argument sequences and presented an analysis and ablation studies showing these improvements are due to our neural model of sequencing and interaction as well as our semantic and structural features, and we conducted experiments showing that we outperform novice humans on the same data, illustrating the difficulty of this task.

One limitation of our approach is that we do not explicitly model structure in the same way as our approach in Section 4.2. We instead capture structure either implicitly with attention mechanisms or using external models of semantic frames and discourse relations. Future work could address this by using approaches such as graph convolutions [Morio and Fujita, 2019] or by jointly modeling persuasion along with argumentative relations, discourse structure, and/or dialogue acts. Likewise, interplay is a simple but effective representation of interaction but modeling threads as dialogues or multi-party discourse rather than monologues may yield further improvements.

Another limitation is that our evaluation used non-expert annotators. Experts in areas such as the psychology of persuasion or experts in domains such as economics or politics may be better attuned to identify persuasive arguments. Furthermore, the users in Change My View are required to provide an explanation for the reason their view changed and we can analyze these reasons and attempt to predict *why* someone changed their view, as in prior work [Atkinson et al., 2019]. We can build on previous work by using *expert rationales* to label persuasive and/or effective arguments.

Future work could build on our work in influence and persuasiveness. This dataset has the advantage of being labeled, but work in *unsupervised* persuasiveness prediction, given only text responses indicating persuasiveness, is one possible direction. This work would build on previous work in agreement and disagreement detection, where explicit phrases such as *"Good arguments"* and *"You haven't really convinced me"* could be identified as markers of persuasion for distant-labeling. In the same vein, other proxies like community voting [Wei et al., 2016] or argument quality [Wachsmuth et al., 2017b] and convincingness [Habernal and Gurevych, 2016a] could be explored in terms of their correlation to personalized persuasion. Arguments with a community consensus for these aspects are also likely to be persuasive to individuals, conditioned upon other factors such as prior beliefs [Durmus and Cardie, 2018a] and personality [Lukin et al., 2017].

Chapter 5: Controlled Generation of Argumentative Content

In Chapter 4, we discussed the properties of arguments in terms of their structure and semantics as well as some aspects of what makes an argument persuasive. One aspect of persuasion is the provision of novel content – one is more likely to be convinced of a counter-argument if it contains information of which they were not previously aware [Tan et al., 2016, Wachsmuth et al., 2018b]. However, while novelty is one aspect that we can incorporate into models of argumentation and persuasion, questions remain as far as how to *provide* novel content in a counter-argument. In this chapter, we discuss how aspects of argumentation such as novelty and structure can be leveraged to provide effective arguments.

Another aspect of argumentation, not addressed in Chapter 4, is the role of stance. Previously, in our analysis of arguments, we were given counter-arguments with the opposite stance. However, in a generation setting, the stance of an argument must be either predicted or determined during generation. We thus present complementary approaches for *retrieving* a counter-argument with a contrasting stance and *generating* a counter-claim with a contrasting stance.

In Chapter 2.2, we discussed prior work in *retrieval-based* and *generation-based* argumentation. *Retrieval-based* approaches have the advantage of being fluent and globally coherent but the downside is they are inflexible and may not directly address the topic or the main claim of an argument. *Generation-based* approaches on the other hand can adapt their output to the topic or main claim, but may struggle with semantics and global coherence. This distinction can be compared to that of extractive and abstractive summarization [Chen and Bansal, 2018], where each approach has specific advantages.

We present a complementary set of approaches to counter-argument generation – a *hybrid retrieval/generation-based* approach, which identifies an appropriate argumentative response and makes minor functional modifications, and a fully *generation-based* one, which creates argumentative content by making semantic substitutions. For the first approach, we discuss in Section 5.1 how we retrieve counter-arguments using state-of-the-art methods for detecting contrast and our methods for persuasion detection. We then modify the retrieved arguments to select the appropriate connective between two segments while keeping the content the same. In comparison, we also present an approach in Section 5.2 to modify the content of a claim. In this generation-based approach, we make semantic edits to change the claim to be contrastive with the original version [Hidey and McKeown, 2019].

The advantage of argument generation is that it allows us to study these aspects of argumentation in a controlled environment. The controlled environment allows us to determine the impact of different aspects of persuasion, such as the role of argument structure. This environment also allows us to conduct our experiments without being overly concerned about adverse effects that would occur due to deploying a chatbot that discusses issues on the Internet (as discussed in Section 1.2). We thus build models for generating arguments and evaluate them using human annotators.

5.1 Hybrid Retrieval and Generation of Counter-Arguments

In order to make progress towards effective argument generation, an ideal approach would incorporate the aspects of persuasive arguments discussed in Sections 4.1 and 4.3 as well as in prior work. These aspects include stance correctness [Sobhani et al., 2015], semantic coherence [Hidey et al., 2017], argument structure [Ghosh et al., 2016], and the use of discourse relations [Durmus et al., 2019b] and dialogue acts [Yang et al., 2019]. While neural methods for language generation have improved significantly in recent years, both for open-ended generation [Radford et al., 2019] and tasks such as summarization [Lewis et al., 2019] and argumentation [Hua and Wang, 2018, Hua et al., 2019a, Hua and Wang, 2019], these methods have been shown to have a number of limitations. In particular, neural methods for machine translation have been shown to alter the meaning of an input sentence [Vyas et al., 2018] and neural methods for summarization have been shown to hallucinate incorrect information [Kryscinski et al., 2019]. They have also been shown to have limitations in the generation of structure if the structure is not explicitly modeled

[Reed et al., 2018, See et al., 2019]. Furthermore, given the lack of progress in the *identification* of discourse relations [Roze et al., 2019], *generating* consecutive sentences with correct implicit semantics is unlikely without explicit training.

On the other hand, neural models have been shown to be able to generate generic phrases such as *"I don't know"* or *"I think this is true"* [Li et al., 2016a], which have a key rhetorical function in argumentation. Our experiments in Section 4.1.3 showed that agreement and disagreement were key to persuasion.

We thus create a *hybrid* model for counter-argument *fusion* that incorporates both *retrieval-based* and *generation-based* components. The *retrieval-based* components allow us to identify counter-arguments that have the correct stance, components of argumentation, and aspects of persuasion. The *generation-based* component allows us to *fuse* counter-arguments by fluently transitioning between argumentative segments and by generating explicit discourse relations and dialogue acts.

Previous work has mostly focused on combining claims and premises [Rinott et al., 2015] or on generating short paragraphs [Hua and Wang, 2019]. However, many arguments consist of multiple claims that respond to a main claim and in turn have supporting arguments. Thus, a key aspect of argument generation is to combine multiple related argumentative "sub-trees." One approach would be to identify the argumentative structure and then consider these sub-trees as candidates for argument fusion. However, our work in Section 4.2.3 showed that this is a difficult task in internet dialogues and we instead use paragraph boundaries, which are likely to contain entire arguments and have been used for training abstractive generation-based models in prior work [Hua and Wang, 2018]. Our overall experiments are thus conducted at the *paragraph* level, whereby we identify candidate paragraphs and then fuse them together by editing argumentative "shells," generic non-content phrases with a rhetorical function [Madnani et al., 2012] such as *"Good idea"* or *"For my second point."* As we consider already-formed arguments, our work is agnostic to the method of argument creation and is thus complementary to existing work. While we use retrieved and identified paragraphs, we could use arguments created by retrieving entire arguments [Wachsmuth et al., 2018b], synthesizing claims and premises [Bilu and Slonim, 2016, El Baff et al., 2019], or generating entire arguments abstractively [Hua and Wang, 2018, 2019].

Our approach consists of three primary components to address the aforementioned issues:

1. **Stance Detection:** We create a dataset of pairs of discussion threads from Change My View labeled for stance and use these pairs to train a model for stance prediction.

2. **Persuasive Argument Extraction:** We build on our work in Sections 4.2 and 4.3 to identify the paragraphs in an argument that contain argumentative content and are likely to be persuasive.

3. **Fusion:** We train a neural generative model to replace argumentative shells between paragraphs, allowing us to model discourse and dialogue coherence.

In Section 5.1.1, we discuss how we partitioned the data from Change My View to train each subcomponent- **labeling a new subset of data for stance prediction**, leveraging the existing data for persuasive argument detection, and setting aside the remainder of the data for argument fusion. In Section 5.1.2, we discuss our approaches for each of the three sub-tasks, using a state-of-the-art approach for relation prediction [Devlin et al., 2019] to select counter-argument candidates with the correct stance, **building on our approach for persuasive argument detection** from Section 4.3 to select and order paragraphs from the counter-argument candidates, and finally using a state-of-the-art approach for text-to-text generation [Lewis et al., 2019] to edit the paragraphs by replacing argumentative shells, e.g. explicit discourse and dialogue connective phrases. Finally, we present an intrinsic analysis of each of the three components in Section 5.1.3 and an overall evaluation of the resulting counter-arguments using human annotators (Section 5.1.4).

5.1.1 Data

In order to create models for each sub-task, as with our work in Chapter 4 we again use Change My View for our experiments. As Change My View is a continually growing resource, we collect additional data to use for training and evaluation. Using the online collection of Reddit comments from pushshift,[1] we obtain discussion threads from Change My View between 2013/01/17 and 2019/08/14. In total, this results in 19,609 discussion threads. We partition these discussion threads into three subsets: 1) a dataset for stance detection 2) a dataset for persuasive argument detection and 3) a dataset for argument fusion.

Stance Detection One key aspect of counter-argument generation is to ensure that the counter-argument has the correct stance. While we could learn stance as part of the end-to-end process during generation, instead we follow previous work and retrieve counter-arguments due to the limitations of generation [Wachsmuth et al., 2018b, Stab et al., 2018b]. Although prior work has resulted in the creation of stance-labeled datasets for argument retrieval, our motivation for creating a new dataset is three-fold:

1. Many argument datasets are labeled for an argument in relation to a topic rather than an argument [Wachsmuth et al., 2018b, Stab et al., 2018b] and are limited to a fixed number of common topics such as abortion, gun control, gay marriage [Napoles et al., 2017]. Instead, we label counter-arguments in response to a main claim and argument and we do not limit the topic domain.

2. Many argument datasets are heterogeneous, with arguments obtained from free text from Google search [Stab et al., 2018b], or parliamentary proceedings and Wikipedia articles [Bar-Haim et al., 2017a], for example. While heterogeneity is the ideal scenario for argument retrieval, our focus on arguments from a single source (Change My View) makes it easier to conduct controlled experiments on argument fusion with multiple authors rather than to also handle domain adaptation.

3. Change My View has several properties which, due to its nature, result in a more structured discussion than we would find in other sources. For example, the rules of Change My View (discussed in Section 3.1.1) require that all responses to an original post are a refutation, which means that if we can determine the relation between two original posts, we can assume

[1] http://pushshift.io

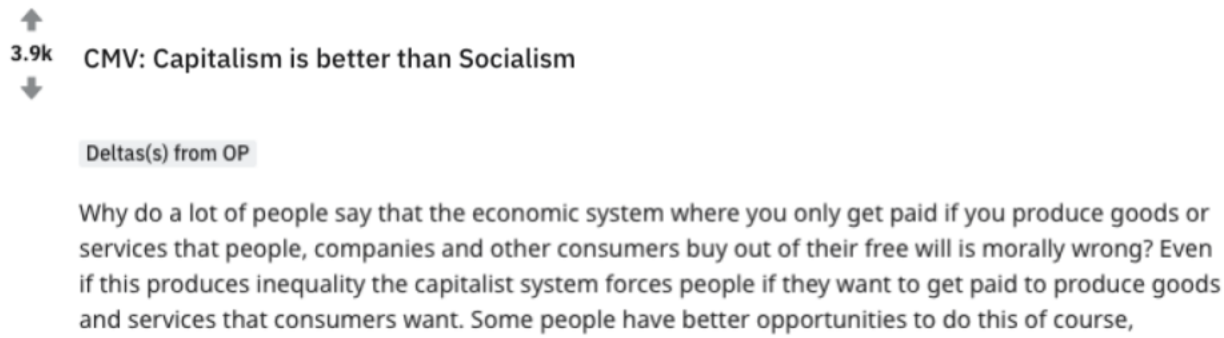


Figure 5.1: An example title and truncated original post from Change My View.

that the responses are all counter-arguments. Additionally, we can again leverage the fact that Change My View is naturally-labeled for persuasion. While datasets for social media exist such as the SemEval 2016 Stance Detection for Twitter task [Mohammad et al., 2016] and Interpretable Semantic Similarity task [Agirre et al., 2016], these datasets do not also provide persuasive supporting counter-arguments.

Given these desired properties, we make the assumption that many arguments on Change My View have been discussed previously.[2] In order to obtain candidate counter-arguments for stance detection, we use Google search, which has been used for other natural language processing tasks such as fact checking [Mihaylova et al., 2018, Chakrabarty et al., 2018] or argument retrieval [Stab et al., 2018b]. Consider the example in Figure 5.1. If we use Google to search for the title[3] of the post *"Capitalism is better than socialism,"* we obtain the results shown in Figure 5.2. On this particular topic, we obtain a number of relevant discussion threads, with arguments both for and against the title claim of the **"provided thread,"** or the claim we wish to identify counter-arguments for.

In our scenario, our goal is to create a counter-argument provided a **title** (the main claim) and an original post (**OP**) in support of the title. During training, we have access to the entire

[2] In fact, the moderators might agree with this assertion as the rules for "Fresh Topic Friday" require that new submissions be on topics that were not discussed in the month previous, suggesting that re-submissions are a common occurrence.

[3] which is also the main claim, as discussed in our annotation of Change My View in Section 4.1

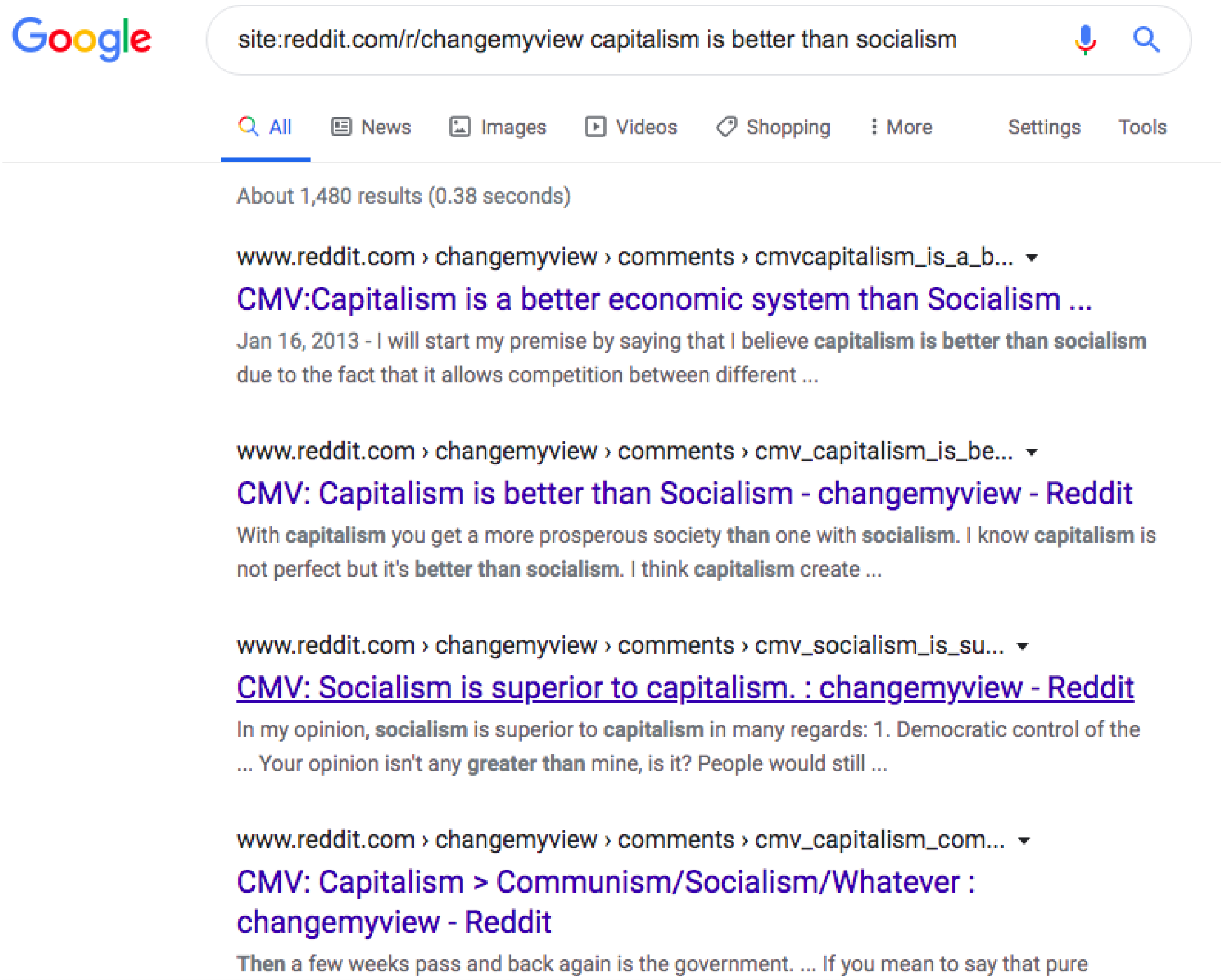


Figure 5.2: The top-ranked Google search results of the title from 5.1

provided discussion **thread**, including **response** posts, which by the nature of CMV are always counter-arguments to the OP. During inference, we only have access to the OP as our goal is counter-argument generation– if we had access to the responses in the thread we could simply provide one of the responses. However, during both training and inference, we also have access to a set of **retrieved threads** from Google search, each of which consist of a title and OP along with a set of responses. Then, given a relation (AGREEMENT or DISAGREEMENT) between the provided title and a retrieved title, we can make an assumption about whether the retrieved original post or responses either SUPPORT or ATTACK the provided title[4] and can act as counter-arguments if the latter. In other words, if the titles of a provided and retrieved thread have the *same* stance, the responses in the retrieved thread will be counter-arguments and have the *opposite* stance of the provided title, by nature of the rules of Change My View. Alternatively, if the titles of a provided and retrieved thread have *opposite* stances, the retrieved OP, which supports the title, can be assumed to be a counter-argument. This assumption is diagrammed and exemplified in Figure 5.3. In this example, the title from the provided thread *"Capitalism is better than socialism"* and the title from one of the retrieved threads *"Socialism is superior to capitalism"* have opposing stances. We can then use the retrieved OP as a counter-argument. If the provided and retrieved examples had the same stance (e.g. if the provided claim was instead *"Capitalism is the worst economic system"*), we could instead use all of the responses from the retrieved thread, as the rules of CMV require these to be counter-arguments.

However, Google search only provides the most relevant documents for a search query and does not indicate the stance of a counter-argument in relation to a claim. Thus, we conduct an annotation experiment on Amazon Mechanical Turk (AMT) to obtain stance labels for the relation between two discussion threads – AGREEMENT, DISAGREEMENT, or UNRELATED. We ask AMT workers only to consider the relation between the two main claims of the post, as per the rules of Change My View, the titles must state the point of view of the poster (Section 3.1.1) and we therefore do not need to resolve co-reference and other anaphora. This setting makes the task similar to

[4]to use the terminology from Section 4.1.1

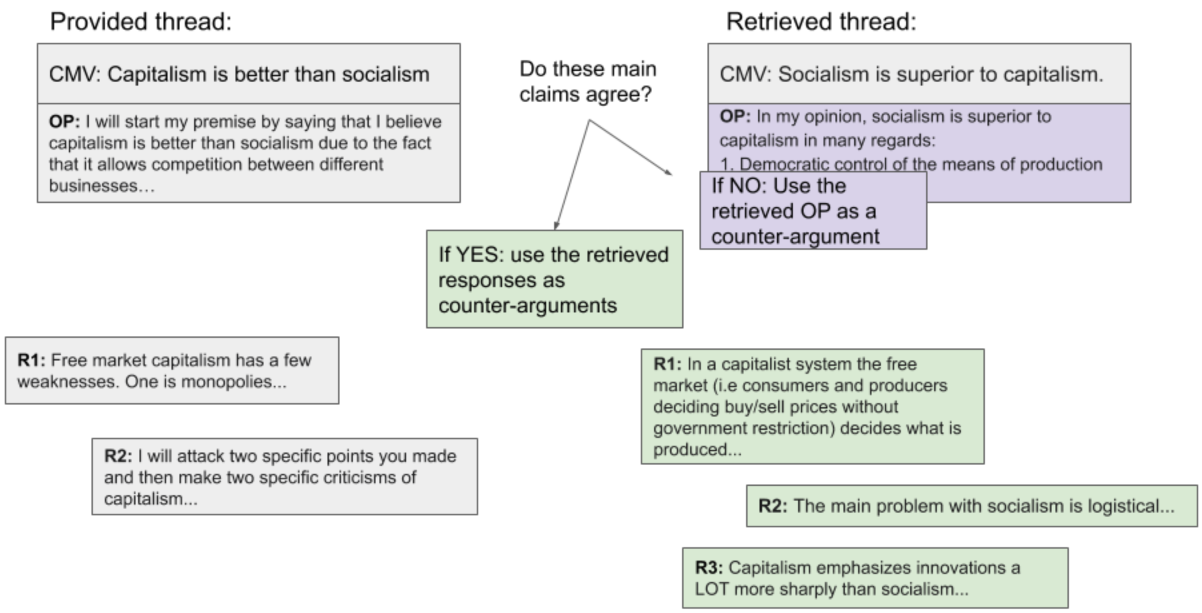


Figure 5.3: The truncated discussion thread from Figure 5.1 and one of the retrieved threads from Figure 5.2. Since the title claims disagree, we can assume that the provided original post disagrees with the retrieved title and that the retrieved original post disagrees with the provided title and we can use the retrieved original post as a counter-argument. If the title claims agreed, we could use the retrieved responses (R1, R2, and R3) as counter-arguments to the provided title claim.

natural language inference [Bowman et al., 2015], among other tasks, where crowd workers have demonstrated the ability to annotate short texts. We thus assume that AMT workers are capable of this task as our framing of the task only requires annotating two sentences with a relation rather than entire discussion threads.

First, we set aside discussion threads from our collected CMV data between the dates of 2019/04/10 to 2019/08/13, which we subject to a number of restrictions. We noticed that the data obtained from Pushshift, which is a snapshot of Reddit at a certain point in time, contains several threads that were later removed by moderators. Upon close examination we find that many of the removed posts contain explicit language that could be considered hate speech. Thus, in addition to using the Reddit API[5] to ensure that our retrieved posts have not been removed by moderators, we take a conservative approach to content selection by removing all posts from the dataset that contain keywords related to specific topics. This results in the removal of any topics associated with religion, race, ethnicity, and gender, leaving 1,814 discussion threads (the set of "provided threads") whose claims we submit to a Google search.

We used the Google search API,[6] which allows up to 10 search results, and we limited the site domain to `https://reddit.com/r/changemyview`. As Google is a black box, we apply additional post-processing to the 10 retrieved discussion threads to determine whether the thread should be annotated. We again filter hate speech using the same parameters as for our provided threads. We also require that the retrieved threads have not been removed or deleted and that the date of the post is earlier than 2019/4/10, as this allows us to simulate a real-world scenario, where we cannot retrieve arguments from the future. We also discard threads that have the exact same main claim as the claim from the provided thread. This process resulted in 10,555 provided/retrieved thread pairs, for an average of 5.82 retrieved threads for the 1,814 provided threads.

For each of the 10,555 thread pairs we extracted the title/main claim of each thread and removed any reference to *"CMV."* Then, we asked annotators to *"judge the pair of opinions based on their*

[5] `https://praw.readthedocs.io/en/latest/`
[6] `https://developers.google.com/custom-search/v1/overview`

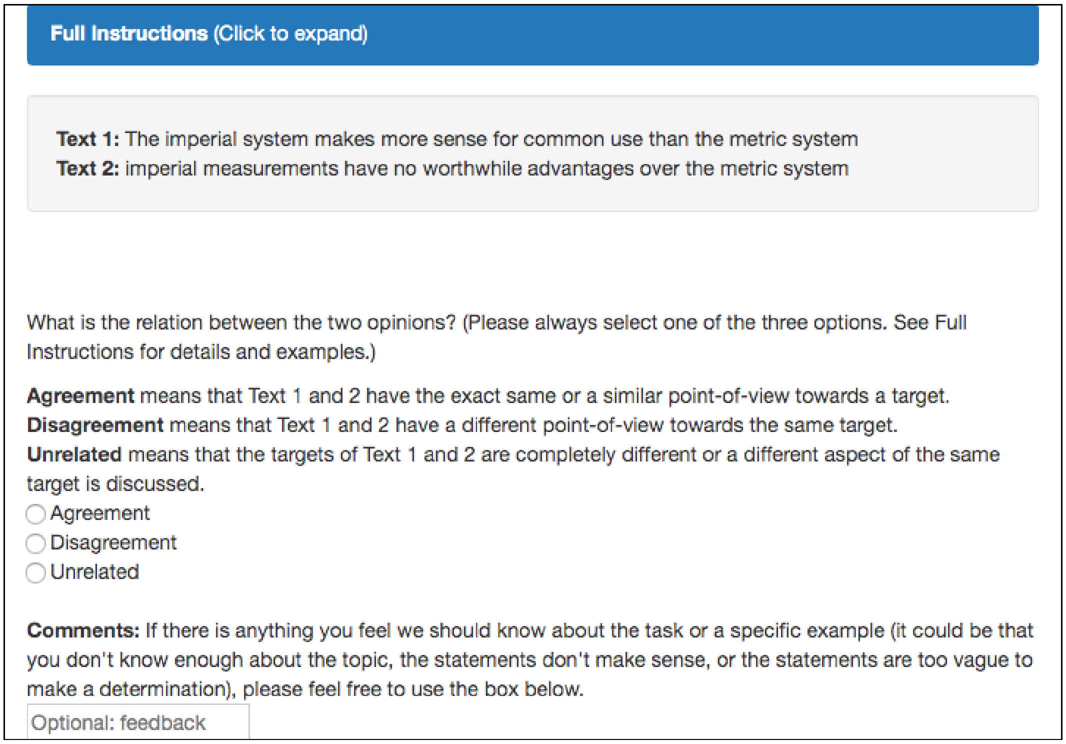


Figure 5.4: An example task from the interface presented to AMT workers.

similarity: whether the pairs agree, disagree, or are unrelated. To label the pairs as agreement or disagreement, they should have the same target, or in other words should have the same/opposing opinion in relation to the same subject. Otherwise, they should be labeled as unrelated." We also instructed them that if the pairs partially agree or partially disagree, they should be labeled as AGREEMENT or DISAGREEMENT, respectively, and if the pairs are discussing a similar topic but a different aspect of that topic, label them as UNRELATED. To further illuminate the instructions, we provided the example pairs in Table 5.1, along with an explanation. We also provide an example from the task in Figure 5.4. The "Full Instructions" include the description above along with the full set of examples from Table 5.1.

Before conducting our annotation experiment, the task was approved by the Columbia Institutional Review Board (IRB-AAAS1977). We paid annotators at the rate of the US federal minimum wage, \$0.05 per example under the assumption that each example would take around 20 seconds. We required that workers are from an English-speaking country (USA, Canada, UK, Australia, or

Text 1	Text 2	Label	Reason
Capitalism is a better economic system than Socialism	Capitalism is the best economic system	Agreement	Both Text 1 and Text 2 have the same opinion of the target, capitalism.
Capitalism is a better economic system than Socialism	Socialism is superior to capitalism	Disagreement	Text 1 and Text 2 have different opinions of the target, capitalism
Automatic Software Updates Should Be Illegal and Require Consumer Consent.	All two-party consent laws should be repealed and replaced with one-party consent laws.	Unrelated	Text 1 and Text 2 have different targets, "automatic software updates" and "two-party consent laws."
the majority of the population won't be correctly represented in politics until there is a major political education reform	I feel that the fundamental problem in American politics is a dangerously ill-informed populace	Agreement	Text 1 and Text 2 have the same target, a lack of political education, represented in Text 1 as "major political education reform" and Text 2 "a dangerously ill-informed populace" and the same opinion toward the target, which is that political education should be increased.
Capitalism is a better economic system than Socialism	I think capitalism as an economic system could be sustainable with proper restraints and a robust socialist government to contain it.	Agreement	Text 1 and Text 2 are both in support of capitalism overall, although Text 2 is only partially in favor, with some qualifications. As this is partial agreement, the correct label is agreement.
the majority of the population won't be correctly represented in politics until there is a major political education reform	Proportional representation is a better system of democracy than single member plurality in almost every way.	Unrelated	Text 1 and Text 2 are both about politics, but have different targets: "political education reform" and "proportional representation," respectively.
Voting Day in the USA should be a National Holiday	Voting Day Should be A Public Holiday	Agreement	While Text 1 and Text 2 have slightly different targets, voting day vs. voting day in the USA, the relation is at least partial agreement, as Text 2 agrees with Text 1 because it is more specific.
Voting Day in the USA should be a National Holiday	Making Election Day in the United States a federal holiday won't make a dramatic change in voter turnout.	Disagreement	The target of Text 1 and Text 2 is the same, election day, and in the case of Text 2 it is implied that they do not believe election day should be a federal holiday.

Table 5.1: The examples and explanations provided to annotators on AMT.

New Zealand) and that they have obtained Master's level and a lifetime approval rating of 99%.

We had 3 annotators for each of the 10,555 pairs. Rather than use a naive voting mechanism, which would result in many ties, we used MACE [Hovy et al., 2013] to determine the "true" label given the annotations. MACE is a Bayesian method of annotation where the assumption is that the true label of an example and the competence of an annotator are both latent variables. MACE models the agreement between annotators as well as the entropy of labels (i.e. an annotator that labels all examples with the same label, in contrast to the distribution of labels, is likely to be a "spammer."). In total, there were 173 annotators and we instructed MACE to remove annotators with quality scores lower than 0.4. This process removed 21 low-quality annotators and the labels were then determined by MACE according to a weighted average of the remaining annotators.

Overall, there are 2,352 examples labeled for AGREEMENT, 1,544 for DISAGREEMENT, and 6,659 for UNRELATED. Of the 1,814 provided threads, 1,217 have at least one retrieved thread labeled for DISAGREEMENT/AGREEMENT, which allows us to use those threads for the end-to-end counter-argument generation task. For our experiments, we perform 10-fold cross-validation for the 1,217 provided threads, setting aside one fold for validation and hyper-parameter tuning and one fold for evaluating our full hybrid counter-argument generation system.

Persuasive Argument Extraction We use the same data for the pairwise prediction task described in Section 4.3.1, removing any retrieved threads that overlap with our stance-labeled data. This results in a nearly-balanced dataset of 6,589 examples for training and 1,516 for testing. There are 3,438 training examples that received a delta, leaving 3,151 unconvincing arguments. Likewise, the numbers are 802 and 714 for testing, respectively.

Fusion We use the remaining Change My View data from prior to 2019/04/10 for training the fusion component, removing any duplicates that overlap with the data used for Persuasive Argument Extraction. This results in 14,176 threads, of which we set aside 2,000 for testing. We select only those responses that received a delta, in order to learn how to generate, for example, explicit speech acts such as agreement that our work has shown to be persuasive (in Section 4.1.4). This results in 49,534 responses, which we split into paragraphs to train the fusion model.

5.1.2 Methods

To output a counter-argument, an ideal approach would only provide arguments that have the opposite stance. The selected arguments would also have the aspects discussed in Chapter 4–coherence and structure, among others. Finally, the arguments would also insert argumentative shells appropriately – transition phrases such as discourse connectives between paragraphs.

To handle these properties, we first **predict the stance** of all posts in the retrieved threads from Google search. Then, we use our model of persuasion and argumentative components to **extract and order the paragraphs** from the posts that are predicted to be counter-arguments. Finally, we **fuse** the extracted paragraphs by replacing discourse connectives and other shell phrases. A schematic of our full hybrid counter-argument system is provided in Figure 5.5. After retrieving full discussion threads using Google search, we identify counter-arguments at the post level (either OPs or responses) using an ensemble of the top-ranked search result and state-of-the-art prediction models for claim relation types (AGREEMENT, DISAGREEMENT, or UNRELATED between pairs of titles) and argument relation types[7] (ATTACK, SUPPORT, or UNRELATED using labeled title and post pairs derived from our claim relation labels). Then, we split the posts into paragraphs, filter those that contain no argumentative content, and score ordered combinations of those paragraphs using a model of persuasive argument detection. Finally, we remove argumentative shells using a rule-based approach and generate new shells using a state-of-the-art model for language generation.

Stance Detection During the stance detection stage, our goal is to identify counter-arguments with the correct stance. As the remainder of the pipeline depends on having a non-null set of counter-arguments, we desire an approach that always provides at least one counter-argument. In other words, we want there to be 100% *coverage* for each provided thread in our dataset. Thus, given the constraint of full coverage, our goal is to maximize the *precision* of the retrieved counter-

[7]to use the terminology from Section 4.1.1

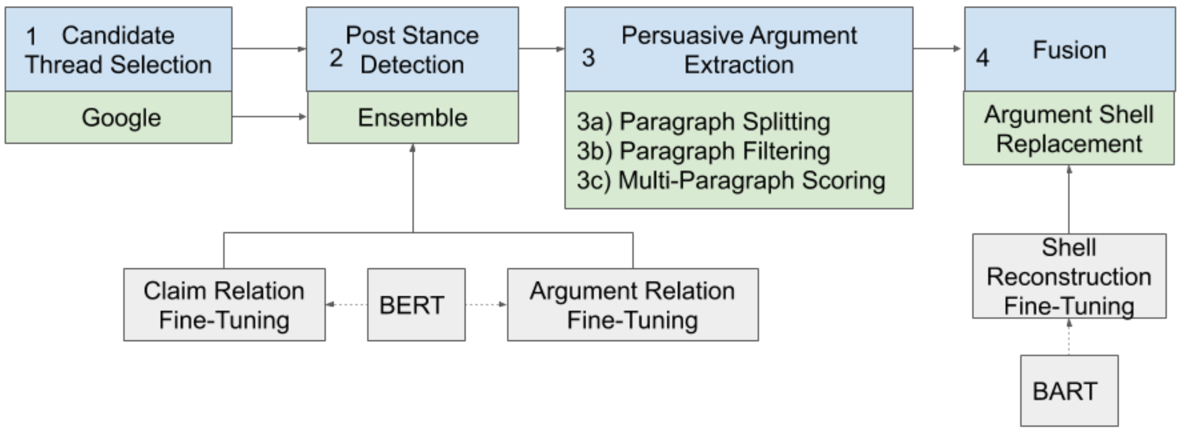


Figure 5.5: Our Hybrid System for counter-argument generation. We first select candidate threads using Google search and predict the stances of individual posts using an ensemble approach. Then we extract individual paragraphs from the posts by identifying argumentative content and scoring the content for persuasiveness. We finally edit the posts to be more coherence using a generation-based approach.

arguments in terms of stance. One possibility is to use the **top-ranked Google** search result, which guarantees full coverage. However, we suspect that this approach would result in low precision on our labeled stance data. Instead, we train two models on our labeled stance data: 1) a model for **claim relation types** that predicts the relation (AGREEMENT, DISAGREEMENT, or UNRELATED) between a provided title claim and a retrieved title claim and 2) a model for **argument relation types**[8] that predicts whether a retrieved post (either an OP or a response) SUPPORTS, ATTACKS, or is UNRELATED to the provided title claim. Finally, in order to maximize *precision* while maintaining *full coverage*, we **ensemble** the three approaches (the two models and the top-ranked Google search result) by first applying the most precise approach and iteratively applying less precise approaches until the set of counter-arguments is non-empty.

Our model for **claim relation types** uses pre-trained contextual embeddings. Specifically, we use BERT [Devlin et al., 2019], as this model has shown excellent performance on related tasks such as natural language inference, where the categories of ENTAILMENT, CONTRADICTION, and NEUTRAL are similar to our categories of AGREEMENT, DISAGREEMENT, or UNRELATED.

[8] to use the terminology from Section 4.1.1

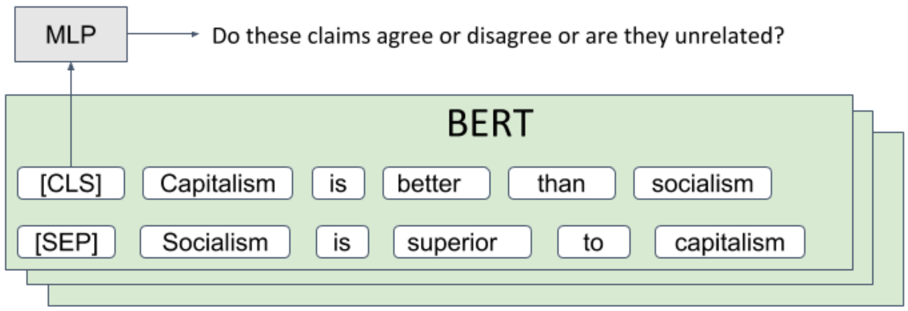


Figure 5.6: A sample training instance for the title pair BERT model.

Likewise, we use BERT for **argument relation types** as BERT performed reasonably well on identifying argumentative relations in Section 4.2.3.

First, to obtain a model for **claim relation types**, we fine-tune BERT on the 10,555 pairs, each consisting of a provided title claim and a retrieved title claim. We use a cross-validation approach where one of the 10 folds is set aside for tuning and one for evaluation. Figure 5.6 depicts a training example and the BERT representation.

Next, we fine-tune a separate BERT model for **argument relation types** on all pairs of title claims and posts. In other words, we expand our labeled dataset to consider not just the relation between a provided title and retrieved title, but also between each title and all four types of posts. During training and inference, we have access to 1) the OP from a retrieved thread 2) all the responses from a retrieved thread and 3) the OP from the provided thread. During training, we also have access to 4) all the responses from the provided thread. Due to the rules of Change My View, we can assume that an OP in a thread is an argument in *support* of a title claim and all responses in a thread are counter-arguments that *attack* the title claim. Given our labeled dataset for stance, we can also assume that the relation between a provided title and a retrieved title allows us to infer the relation between a provided title and a retrieved OP and all responses, and vice versa.

This assumption allows us to build a directed acyclic graph with edges between OPs/responses and titles from different threads. We obtain labels for these edges based on the relation between the provided title and retrieved title. We refer back to Section 4.1.2, where a premise either *supports* or *attacks* another claim or premise. Here, we assume that an argument (an entire post) *supports*,

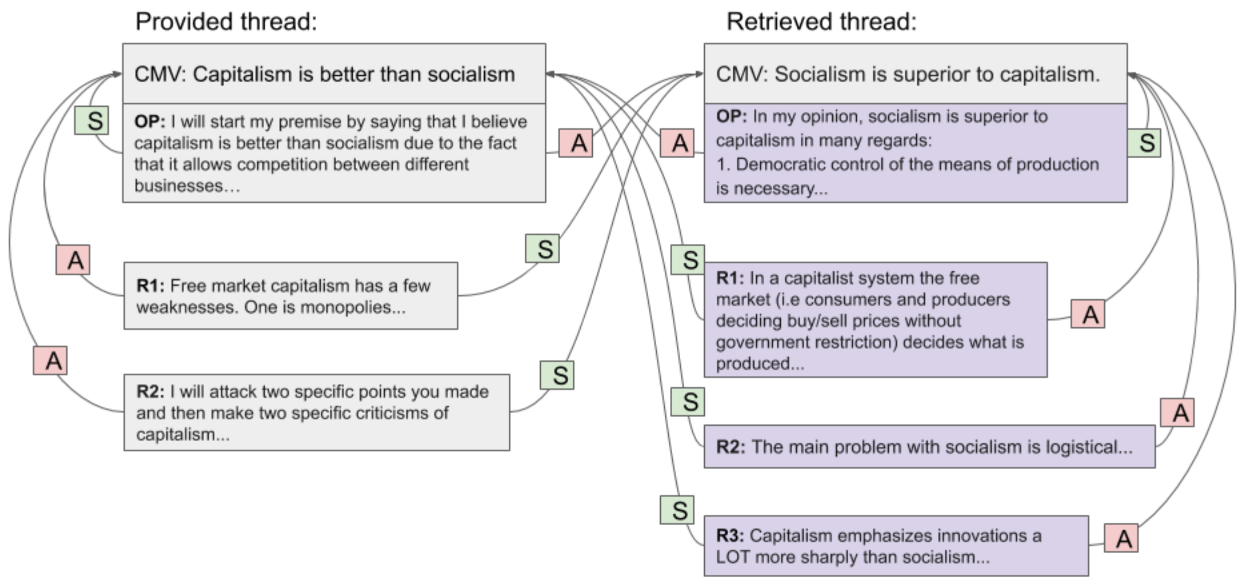


Figure 5.7: An example of a directed acyclic graph obtained from a discussion thread and the Disagreement relation between the provided title and retrieved title. Each edge leads to a title/post training instance, obtained from posts both within and across threads. During evaluation, only the shaded OP and responses are available to the system.

attacks, or is *unrelated* to a main (title) claim. First, an OP in the same thread always *supports* the title and a response always *attacks* the title. Second, the relation between titles and OPs and responses across threads depends on the stance label. Specifically, if the provided title and retrieved title *agree*, then the retrieved OP *supports* the provided title and the provided OP *supports* the retrieved title. In contrast, all the retrieved responses and provided responses *attack* the provided title and retrieved title, respectively. On the other hand, if the provided title and retrieved title *disagree*, then the edge labels would be reversed – OPs *attack* their provided/retrieved counterpart and responses *support* their provided/retrieved counterpart. Finally, if the title claims are *unrelated*, then all edges in the graph are *unrelated*. An example is provided in Figure 5.7. In this example, as the provided title and retrieved title *disagree*, the OPs *attack* the titles across the threads. Additionally, the responses *support* the titles from across the threads.

During training, to maximize the data for modeling, we include all title/OP and title/response

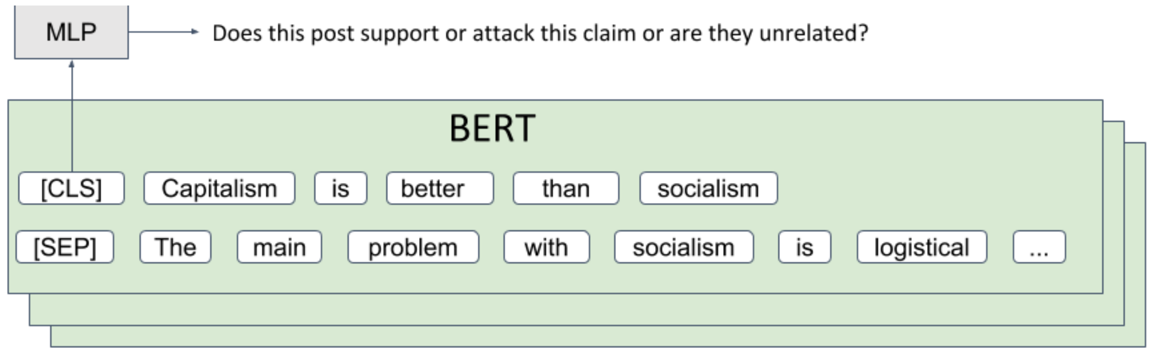


Figure 5.8: A sample training instance for the argument relation type BERT model

pairs, both within and across threads (all the edges from Figure 5.7), resulting in 373,024 pairs. During inference, the candidate pairs are limited only to the provided title claim and OP and responses from a retrieved thread (only the edges from *"Capitalism is better than socialism"* to OP, R1, R2, and R3 of the retrieved thread). This is because in a real-world scenario, only the retrieved threads are available for possible counter-arguments. We fine-tune BERT to predict SUPPORT, ATTACK, or UNRELATED on the title claims and OPs/responses. We train as a separate model from the one for claim relation types because we assume that AGREEMENT/DISAGREEMENT between claims is more closely related to natural language inference, whereas argument relations require more complex reasoning. An example is provided in Figure 5.8. We cross-validate using the same split as the claim relation model.

Our final stance detection system **ensembles** the models for **claim relation types** and **argument relation types** along with the **top-ranked Google** search result to best balance precision and recall. As discussed, one constraint of the system is to have full *coverage* of all arguments, so that candidate counter-arguments can be considered in the next stage of the pipeline. Our end-to-end counter-argument generation system would ideally obtain counter-argument candidates that all have the correct stance. However, the highest-precision approach would result in some cases where no counter-argument candidates are obtained for a given claim, which our system is unable to handle. Thus, we consider a pipeline approach that successively selects counter-argument candidates using increasingly imprecise methods if no candidates were selected at the previous stage.

First, for the highest precision approach, we select retrieved OPs and responses where the claim relation model and the argument relation model have the same predictions of a counter-argument. In other words, we include responses if the predicted claim relation is AGREEMENT and the predicted argument relation is ATTACK and we include OPs if the predicted claim relation is DISAGREEMENT and the predicted argument relation is ATTACK. As we require both models to agree, this stage naturally has higher precision but is likely to result in low coverage. Next, if no candidates are selected at this stage, we consider all retrieved responses where the claim relation model predicts AGREEMENT and retrieved OPs where the claim relation model predicts DISAGREEMENT. Finally, if no retrieved counter-arguments have been identified to this point, we select all responses from the highest-ranked retrieved thread in the Google search results.

This algorithm is outlined as follows:

1. Set of counter-argument posts C = all responses where the argument relation model predicts ATTACK and the claim relation model predicts AGREEMENT
2. If $C = \emptyset$, C = all OPs where the argument relation model predicts ATTACK and the claim relation model predicts DISAGREEMENT
3. If $C = \emptyset$, C = all responses where the claim relation model predicts AGREEMENT
4. If $C = \emptyset$, C = all OPs where the claim relation model predicts DISAGREEMENT
5. If $C = \emptyset$, C = all responses where the retrieved thread was the first Google search result

This approach allows us to obtain counter-argument posts for every provided thread, starting with the highest-precision approach and only using lower-precision approaches if no posts are found.

Persuasive Argument Extraction Given the set of selected post candidates C from the previous step, we *partition* each post to obtain a set of paragraphs P, *filter* the paragraphs for argumentative content using our model from Section 4.2.2, and *score* them using a modified version of our approach to persuasive influence detection from Section 4.3.2. We make the assumption that each paragraph is a complete argument, as in prior work [Hua and Wang, 2018]. Our first step is then to *split* the posts into paragraphs whenever two consecutive newlines appear. To increase the likelihood that each paragraph is itself an argument, we add a *filtering* step, where we remove any paragraphs that contain no claims or premises (using our model from Section 4.2.2) or are less than M characters in length. To *score* the remaining paragraphs, we train a model for persuasive argument detection using the data discussed in Section 5.1.1. Then we apply this model to persuasive argument *extraction* by scoring multiple ordered paragraphs and selecting the K-highest scoring ones. We use a beam search for up to T paragraphs, combining paragraphs and re-scoring them. The final highest scoring argument at time T is then selected for the next stage.

Our model of persuasive argument detection combines an embedded document representation **h** with global features ϕ (from Equation 4.10 and duplicated here):

$$y = \sigma\left(MLP(\mathbf{h}) + \beta^T \phi\right) \tag{5.1}$$

Our full model of persuasive influence from Section 4.3.2 includes global features based on the interaction of style words and content words and a document representation based on all words, frames, and discourse relations from the original post and response. However, at this stage we are only interested in *content-based* features, as we need to consider every paragraph in isolation. Features based on discourse relations and social interactions such as agreement would not be useful at this stage, as in the following stage we replace the argumentative shell with the most appropriate shell given the context. Thus we instead use a set of global features based only on *content* and only consider *content* words for the embedded document representation.

First, we identify a set of *non-content* words using inverse document frequency, identifying the most common words which are lower than a threshold X:

$$W = \forall w, IDF(w) < X \tag{5.2}$$

Then, we derive a number of features using these non-content words and the content words (the remaining words in the dataset), among other features. For our global content features, we select features based on **length** (commonly used in social media applications [Tan et al., 2016]), **content interplay** (using the global features from Section 4.3.2), **argumentation** (using our model from Section 4.2.2), and the interaction between **paragraphs** for each of these features. Our **length** features are the number of sentences, words, characters, quotes, and URLs. Our **content interplay** features are the same as the interplay features from Section 4.3.2, except using only the interaction between content words - common words, similar fraction in reply, similar fraction in OP, and Jaccard score. Our **argumentation** features are derived from the predicted argumentative components using our model from Section 4.2.2 at the sentence level - whether a sentence is a claim, premise, or is non-argumentative.[9] The derived features include the percentage and number of claims, premises, and non-arguments. We also attempt to capture a measure of structure by including the ratio of claims to premises and binary features for whether there are more claims than premises or whether there are all premises or all claims (as an argument should consist of a claim supported by at least one premise). We derive additional "structure-lite" features by computing the percentage of "supported" claims, i.e. if a claim is preceded or followed by a premise, and the percentage of "supporting" premises, i.e. if a premise is preceded or followed by a claim or premise. Finally, we compute the proportion of the longest "chain" of claims, where a chain is a consecutive sentence of the same type (i.e. an argument that has many claims in a row is unlikely to be a good argument). Each of the **length**, **content interplay**, and **argumentation** features are computed for the entire argument and are also aggregated (e.g. the minimum or maximum) at the **paragraph** level.

Our full set of features is as follows:

1. The set of length features L

2. The content interplay features I

3. The set of argumentation features A

[9] We only use the argumentative component model as the argumentative relation model obtained low precision.

4. The "redundancy" features - the minimum I *between* all paragraphs

5. The minimum, maximum, and mean of L for each paragraph

6. The minimum of I for each paragraph

7. The minimum of the percentage of claims, premises, supported claims, and supporting premises for each paragraph and the percentage of paragraphs with more claims than premises

These features capture the tradeoff between content and length. Longer arguments tend to be more persuasive, as captured by the length features, but we want to avoid gaming the model by just selecting very long arguments. The content interplay and redundancy features capture the novelty of each argument – if there is higher minimum overlap with the original post and the other paragraphs, the argument should be weighted lower. The argumentation features capture a measure of quality – each paragraph and the overall response should be high in argumentative content.

For our document content representation, we use the same approach as in Section 4.3.2, with two minor changes. First, our LSTM operates over paragraphs rather than sentences, as the paragraph is the unit of interest for extraction. Second, from each paragraph we use only the *content* words defined in Equation 5.2. The new model is depicted in Figure 5.9 (which is a modification of Figure 4.19). Our new global features are combined with the modified document representation that considers the content words in each paragraph. We train and validate this model on the dataset discussed in Section 5.1.1.

Next, to find the optimal selection and ordering of arguments, we could score all $\frac{|P|!}{(|P|-T)!}$ multi-paragraph arguments using our full model. However, this approach is prohibitively expensive as we are unable to decompose the problem using an approach such as dynamic programming due to our global features and LSTM component. Instead, we use a beam search to iteratively consider the top K arguments up to length T. We first compute the scores for all arguments of 1 paragraph in length from our candidate paragraphs P, keep the top K, compute the scores for the $K * |P|$ arguments of length 2, keep the top K, and so on. After this stage, we have a single argument of at most T paragraphs.

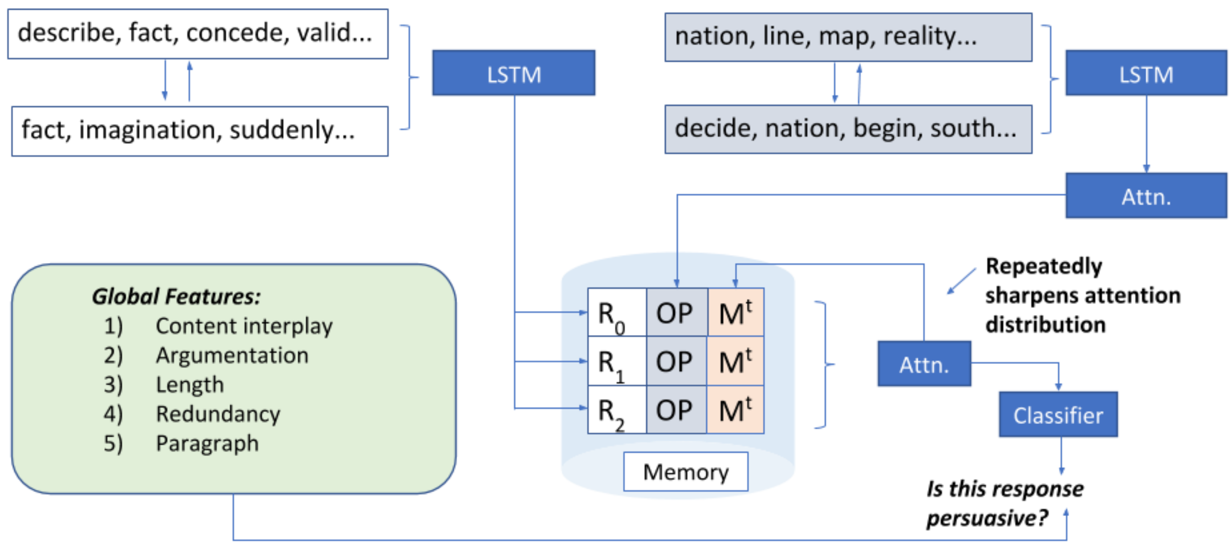
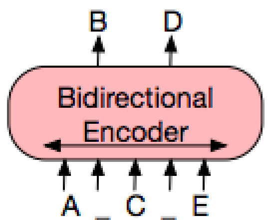


Figure 5.9: A modified version of our persuasive influence network from Figure 4.19 to only use *content* features.

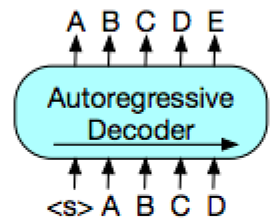
Fusion For the final stage, we edit the resulting argument by *removing* argumentative shells between paragraphs using a rule-based approach and *inserting* new shells using a fine-tuned sequence-to-sequence model. Specifically, we use BART, a large pre-trained sequence-to-sequence model that has obtained state-of-the-art performance on many generation tasks such as machine translation or summarization [Lewis et al., 2019]. As pre-trained language models have been particularly adept at generating fluent output [Radford et al., 2019], we hypothesize that this model will be able to generate fluent transition phrases between paragraphs. A comparison of BART to other approaches is presented in Figure 5.10.[10] Unlike BERT, which was trained using a masked language modeling objective [Devlin et al., 2019], BART is capable of generating output of arbitrary length and is thus more suitable for this task. GPT-2 [Radford et al., 2019], on the other hand, is capable of generating text of arbitrary length, but our task requires the reconstruction of deleted elements, which BART has been trained to do. Furthermore, GPT-2 is conditioned on previously generated text, whereas in our case we want to condition on a number of different contexts such as the original post and left and right sentences.

For the initial component, we use a rule-based approach to identify and remove the shell using

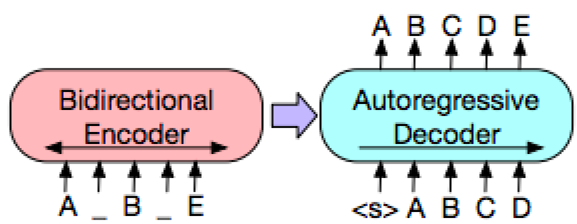
[10]this figure can be found in the work of Lewis et al. [2019]



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 5.10: A comparison of three common pre-training approaches. While BERT uses a masked language modeling approach and GPT-2 allows for language generation of arbitrary length, BART provides the benefits of both.

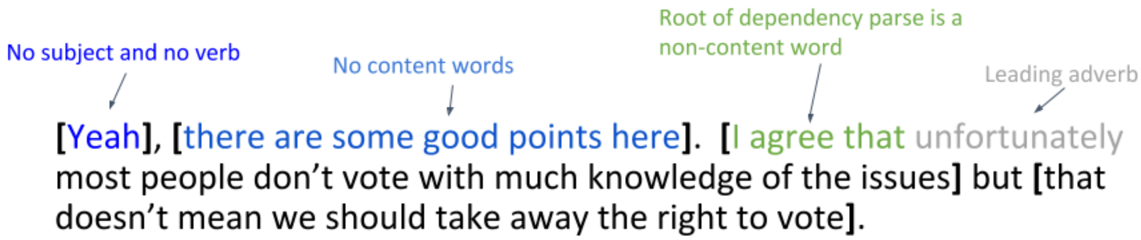


Figure 5.11: An example of the shell identification process. The paragraph is first split into segments according to punctuation and discourse connectives (as shown in brackets). In this example, after successively evaluating each segment for content and identifying initial shell phrases, the argumentative shell is "Yeah, there are some good points here. I agree that unfortunately"

the non-content words from Equation 5.2.[11] First, we sentence tokenize, part-of-speech tag, and dependency parse the paragraph using Spacy.[12] We further segment each sentence by additional punctuation[13] and discourse connectives. We used a binary classifier to identify discourse/non-discourse usage of connectives using the approach of [Braud and Denis, 2016]. For each segment, if there is both a subject and verb, we recursively visit the root node in the dependency parse of the clause until the root node is a content word. This allows us to remove phrases such as *"I agree that."* If no content is identified, we remove the current segment and visit the next one. Finally, we remove any remaining segment-initial adverbs or subordinating conjunctions. An example is provided in Figure 5.11. After this process, the identified shell consists of all the removed segments and initial words in the final segment, which we use as target data during training and which we replace with the predicted shell during inference. While this approach may remove some content, as the second component inserts new shells, it is more important to remove unnecessary jarring transition phrases and so we prefer a higher-recall approach that may result in lower precision. We select as training data only those paragraphs that the rule-based approach identified as having an argumentative shell.

For the second component, we train BART on the selected paragraphs to recover the argumen-

[11] Qualitatively, we observed that the non-content words also include many functional terms such as *"think"* or *"point"* and thus used the same subset as the words in our persuasive argument extraction.

[12] `https://spacy.io`

[13] characters that are likely to separate propositions – ",;;-(){}[]"

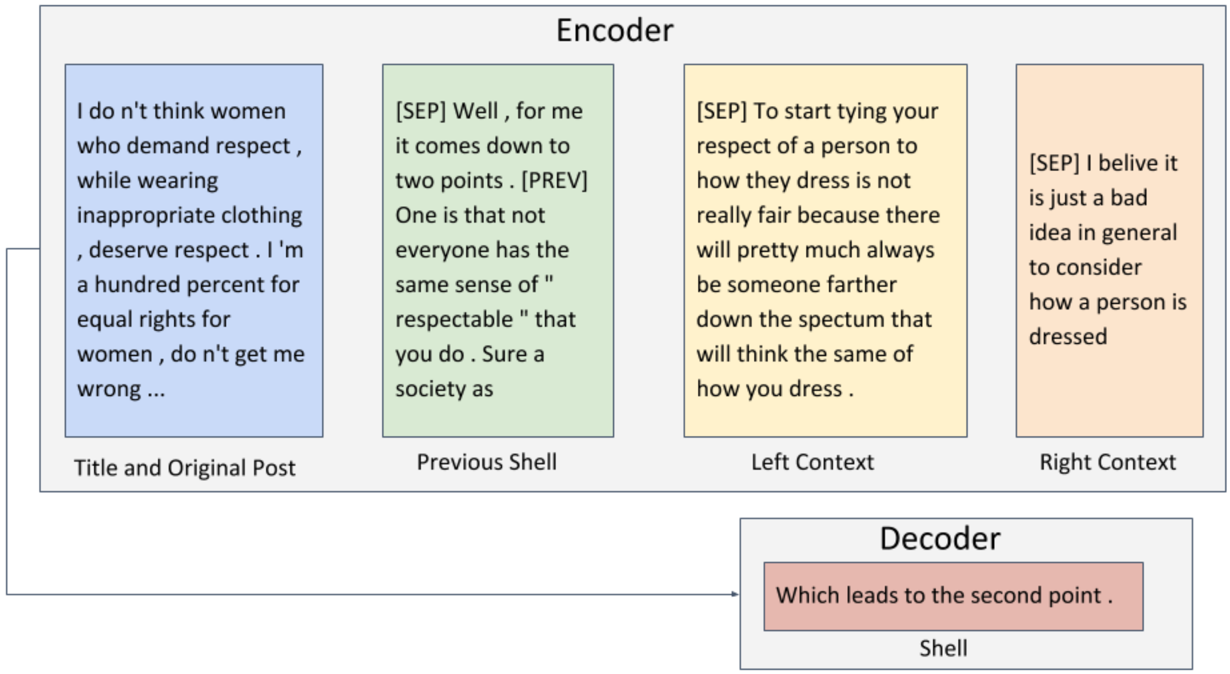


Figure 5.12: An example of the input to the BART encoder and desired output for the decoder to generate. The title and original post, previous shell (from earlier paragraphs in the same post), left context (last sentence in the previous paragraph before the shell) and right context (remainder of the current sentence containing the shell or the next sentence after the shell) have their tokens concatenated along with the special token "[SEP]." The shell, which occurs in between the left and right contexts and may be any number of complete sentences followed by a partial sentence, is the target text for generation.

tative shell given the context. For context, we use the title and first few sentences of the original post, the last sentence in the previous paragraph, and the remainder of the sentence or next sentence following the removed shell. We also include the previous shell, as this allows the model to learn to use dependent phrases such as *"First ... Second"* or *"1) ... 2)"* that are common in multi-paragraph arguments.[14] An example is provided in Figure 5.12. We train BART using the fusion dataset described in Section 5.1.1.

[14]While a hierarchical sequence-to-sequence approach that considers the previously generated shell may improve performance, we consider that for future work.

5.1.3 Experiments and Results

We present an intrinsic evaluation of each of the three components individually (Section 5.1.4 conducts an overall evaluation). First, we report the results of stance detection on the held out test data. Then, we discuss the performance of our persuasive influence detection model on the Change My View test set.[15] Finally, we report the performance of our fusion approach in terms of reconstruction of the argumentative shell in our paragraph-level data as we do not have the true argument shell for our test set. For each component, we also present an ablation study to illustrate the impact of our design choices.

Stance Detection We use version 0.5.0 of the Huggingface library[16] to fine-tune the "BERT-base" model on title pairs and title/post pairs using the default settings. We lowercase all tokens and use the default BERT tokenizer. As recommended by Devlin et al. [2019], we select hyper-parameters by grid search on the validation set over 16 and 32 for batch size, 2e-5, 3e-5, and 5e-5 for learning rate, resulting in a batch size of 16 and learning rate of 5e-5 as the best parameters for both models. We also find that training for 5 epochs (beyond the recommended rate of 4) obtained the best results for both models.

We present our results at the *post-level*, rather than the *thread-level*, as we are interested in the precision and recall of each counter-argument candidate. At the post-level, we evaluate for each post from the retrieved thread (OP and all responses) whether we accurately predict if the post *supports* or *attacks* the provided title, according to our labels derived in Section 5.1.2. We present precision, recall, and F-score for the SUPPORT and ATTACK classes[17] for the **claim relation type** model, the **argument relation type** model, the **top-ranked Google** search result approach, and the **ensemble** of the three approaches. The argument relation type model provides us with a prediction for every provided title and retrieved OP/response pair, but the claim relation type model and

[15] Ideally, we would have labeled data for extracting the most persuasive paragraphs. Instead we evaluate our approach in Section 5.1.4 using human annotators.

[16] `https://github.com/huggingface/pytorch-pretrained-BERT`

[17] We do not include the results for the Unrelated class, as our goal is to retrieve candidate arguments and higher performance on this class does not help with this goal.

	Attack			Support			Thread-Level		
Model	P	R	F	P	R	F	Cov.	#Post	#Par.
Claim Rel.	64.3	46.9	54.3	36.9	44.5	40.4	87.4	13.8	49.2
Arg. Rel.	48.3	58.1	52.7	32.2	24.4	27.8	98.4	21.6	70.8
Google	37.5	54.6	44.4	0	0	0	100	24.7	77.2
Ensemble	63.3	35.7	45.7	41.3	10.4	16.6	100	11.1	36.3

Table 5.2: The results of our stance prediction approaches.

top-ranked Google search result do not. The claim relation type model yields a prediction of AGREEMENT, DISAGREEMENT, or UNRELATED at the thread-level between a provided title and retrieved title. The top-ranked Google search result approach is also at the thread level, as it assigns a prediction of AGREEMENT to all threads that are the first search result returned and a prediction of UNRELATED otherwise. Thus, in order to make all models comparable at the post-level, we apply the same labeling scheme that we used to train the argument relation types, but instead use the predicted claim relation type to assign the predicted argument relation type. We present an additional metric at the *thread-level* - "coverage." Coverage is the percentage of provided threads where at least one counter-argument is retrieved and this number should be maximized so that the other components in the pipeline have arguments for selection. Finally, we present the average number of posts per thread (OPs or responses) selected by each method as well as the average number of paragraphs.

We present results for the four approaches in Table 5.2. Individually, the claim relation types model obtains higher precision for attacks than the argument relation types model (64.3% and 48.3%, respectively), which we need for counter-arguments. However, the former only obtains 87.4% coverage and has lower recall, while the latter obtains close to full coverage at 98.4%. In comparison, using the top-ranked search result yields 100% coverage because this approach always provides a counter-argument, but has very low precision (37.5%). The ensemble, on the other hand, obtains full coverage but even better precision than the claim relation types model. At the same time, it provides an average of 11.1 retrieved posts (candidate counter-arguments) per provided thread, only a small decrease from the 13.8 average of the claim relation types model.

Persuasive Argument Extraction We implement our persuasion prediction model in Pytorch [Paszke et al., 2017] using the AllenNLP library [Gardner et al., 2017]. We use binary cross-entropy as the loss function and Adam [Kingma and Ba, 2015] with a mini-batch size of 100 and a learning rate of 0.01. Word embeddings are initialized with pre-trained 300-dimensional GloVe vectors [Pennington et al., 2014]. Out-of-vocabulary words are randomly initialized and fixed during training. We used a bi-directional LSTM with a 100-dimensional hidden state. The attention layers and classification layer used an MLP with a 200-dimensional hidden state and hyperbolic tangent activation function. We stop training after 50 epochs and perform early stopping on a validation set. The document weights β in Equation 5.1 were pre-trained using a logistic regression classifier. For our global features, we select X such that the size of the set of non-content words W is $|W| = 500$.[18] For extraction, we set the minimum character length M of a paragraph to 30, the number of paragraphs T to 3, and the beam width K to 5.

We present accuracy and precision, recall, and F-score for the positive class in Table 5.3. We compare the full model with the global content features and the document content representation to two baselines. The first consists of only the document representation and the second consists only of the content interplay features used in prior work.

The neural method using only the document content representation obtains the lowest performance on all metrics, suggesting the difficult of modeling argument content without discourse and social interaction features. However, our model of global features, which includes argument structure, obtains a large improvement over the neural approach alone. Combining these two models obtains the best performance in terms of accuracy, recall, and F-score.

Fusion To evaluate the performance of the fusion model, we use a discourse parser to identify the explicit discourse relations in each post. We use the approach of Braud and Denis [2016] to identify discourse usage and spans and the approach of Varia et al. [2019] to predict explicit discourse relations in the 15-way setting. Then, we evaluate how well our approach can *reconstruct*

[18] We experimented with different values but found this to be the best trade-off qualitatively between coverage of non-content words and exclusion of content words.

Model	Accuracy	Precision	Recall	F-score
Content Representation	56.08	57.31	47.64	52.03
Global Features	59.83	63.88	55.36	59.32
Content Interplay	57.12	60.86	53.12	56.72
Full	60.42	60.37	59.55	59.96

Table 5.3: The results of our persuasive argument prediction approaches on the heldout set. The full model is the model from Figure 5.9. Content Representation refers to only using the neural document representation $\mathbf{h}$. Global Features refers to only using the features ϕ. Content Interplay refers to only using the content interplay features I but not the paragraph features.

the same discourse relations when the shell containing the explicit discourse connective is removed and replaced with a new shell. We again use the model of Varia et al. [2019] to predict the explicit discourse relations in the reconstructed shell and evaluate how often the relation in the reconstruction matches the original relation. As a baseline, we compare to a BERT model trained directly to recover the explicit relations given the post with the shell removed. While we would ideally have class labels for all types of shells, the task is open-ended so we instead focus on discourse relations, which have a defined set of classes. Furthermore, we limit our evaluation to include only those examples where the fusion model inserted shell with an explicit discourse connective. Many other types of discourse connectives exist (e.g. the open-ended class of alternative lexicalizations [Prasad et al., 2010]), which would mean the fusion model maintains the discourse relation while using a different realization. However, we do not have a reliable way of evaluating this scenario, so we only consider the case where the model switched to an incorrect relation.

We fine-tune BART using version 0.9.0 of Fairseq [Ott et al., 2019]. We set aside 6,253 examples for evaluation and use the remaining 43,281 examples for training and validation. Of the 6,253 test examples, 2,069 contained an explicit discourse connective. We fine-tune for a maximum of 150 epochs using the default parameters and early stopping on the validation set. We fine-tune BERT using the same settings as stance detection, with the hyper-parameters for batch size, learning rate, and number of epochs set to 16, 5e-5, and 3, respectively.

We report the results of this evaluation in Table 5.4. We use accuracy and weighted precision, recall, and F-score (because the class distribution is very unbalanced). The BART fusion approach

Model	Accuracy	Wtd. Precision	Wtd. Recall	Wtd. F-score
BERT	39.1	30.1	39.1	30.6
BART	35.4	37.0	35.4	35.5

Table 5.4: The results of our BART fusion model on recovering discourse connectives compared to BERT.

obtains comparable accuracy but better weighted F-score than BERT due to better weighted precision. Upon examination of the data, we find that BERT never predicts certain classes such as instantiation, whereas BART tends to insert *"for example"* as a realization of this class. In contrast, BERT over-predicts conditional statements (e.g. *"if"*). We also find that both models make similar predictions for inserting conjunctions (e.g. *"also"* and *"and"*) whereas BART inserts temporal and causal connectives (e.g. *"when"* and *"because"*, respectively, and BERT almost never predicts these classes. BART may have the advantage of being able to learn from similar connectives that are not in the class of explicit discourse markers and is thus able to generalize to specific classes better.

Furthermore, the fusion approach provides two additional advantages: 1) providing the argumentative shell and 2) handling open-ended classes. Using BERT to predict the class of connective phrases would then have two issues. First, we would need to determine what types of relations to handle– here we only considered discourse relations – but in practice the connective phrases we could use are open-ended and we would need to handle speech acts and other types. It would also require realizing the relation via the argumentative shell, where we might use a BART model or another generative model regardless.

5.1.4 Analysis

We conduct an additional experiment to compare our model to an end-to-end state-of-the-art language generation system. Precisely, we train BART [Lewis et al., 2019] to generate full arguments given a title claim and original post (hereafter this model will be referred to as **Baseline-BART**. We use the same dataset we used for training and validating "Our Hybrid System," except we use all responses that received a delta from the original poster, rather than just a subset of paragraphs. In order to train Baseline-BART to generate multiple paragraphs, we use a special paragraph separator '[PAR]'. We train Baseline-BART with the default parameters (as with our fusion model) and select the best model with early stopping on the heldout test set.

We conduct an evaluation of the output using human annotators and then present output from both models to compare them qualitatively.

Human Evaluation We recruit annotators to judge the generated arguments from Our Hybrid System as well as Baseline-BART. Annotators were volunteers from the Columbia Computer Science department and were fluent English speakers, including 5 PhD students, 1 Masters student, and 1 Undergraduate student. We selected these annotators, as opposed to AMT workers, because of the complexity of the task. The annotators were provided with one title claim and original post along with the full output from Our Hybrid System and Baseline-BART.

Annotators were asked to judge each paragraph on four different dimensions: internal consistency, global consistency, transition coherence, and quality. We provided the following definitions to annotators, along with corresponding examples:

- **Internal Consistency** refers to the relations between sentences/clauses within the paragraph. If the paragraph is internally contradictory, it should receive a low score. The examples provided to annotators along with explanations can be found in Table 5.5.

- **Global Consistency** refers to whether the paragraph has the correct stance (i.e. as a counter-argument it should have the opposite stance of the main claim). An argument may score high in internal consistency but not have the correct stance, whereas an internally contradictory argument may have a middling score for stance. The examples provided to annotators along with explanations can be found in Table 5.6.

- **Coherence** refers to the fluency of the transition between paragraphs (e.g. the appropriate use of discourse connectives or speech acts). For example, the first paragraph may start by agreeing with something the original argument said. Additionally, consecutive, related

Example	Main Claim	Counter-Argument	Score
1	Welfare = theft	I agree with you. I don't agree with you.	1
2	Welfare = theft	Welfare is not theft . It 's not theft because it 's the government 's responsibility to help people who are in need . It is theft because the government has a duty to help those in need , not because they 're stealing from the people who need it .	2
3	The United States should not cancel student loan debt	You 're a metaphorical gun pointed to your head by many . It 's a " either you 'll flip burgers all your life for 7$/hour or you accept to be indebted a scandalous amount of money to use " . As such , cancelling student debt is deemed as acceptable because society as a whole created a situation where you were forced to accept a scandalous debt .	5

Table 5.5: Examples of **Internal Consistency**. For Example 1, if the counter-argument is *"I agree with you. I don't agree with you."*, it would receive a very low score (1). Other arguments may be more complex but still inconsistent. For Example 2 below, it receives a score of 2 because it is somewhat inconsistent - the first two sentences are consistent but the third sentence is not. On the other hand, Example 3 maintains the same position throughout.

paragraphs may have discourse connectives such as *"Also"* whereas unrelated paragraphs may have a transition phrase.

- **Quality** refers to whether "you would recommend the use of this argument, regardless of your own opinion on the topic."[19] In other words, we asked "how well do you think the argument provides reasons in support of the topic?" As this requires more than just stating the opposite of the main claim, generic responses such as *"I disagree"* should score low on quality. The examples provided to annotators along with explanations can be found in Table 5.7.

The annotators were provided with an interface developed using the Amazon Mechanical Turk Sandbox (a free platform for development). An example of the task is presented in Figure 5.13. The annotators evaluated 96 full arguments, or 253 paragraphs for Our Hybrid System and 212

[19]We follow previous work in framing this question this way [Gleize et al., 2019] rather than ask about persuasion, which our experiments in Section 4.3.3 showed is a difficult task for annotators.

Example	Main Claim	Counter-Argument	Score
1	it 's better to have hateful opinions , views and message boards out for the public to see rather then delete them or hide them from reality	I live in the US , and I 've seen a lot of hate speech on the internet . It 's not a good idea to keep it out of the public eye . I 'd rather have it hidden from the public view than to have it out in the open . I think that 's the best thing to do , but there 's a difference between keeping it out there and hiding it from the general public . If you want to keep your opinion out there , you have to be willing to put up with a certain amount of hate , and if you ca n't do that , you need to find a way to hide it .	3
2	Constitutional Monarchy is the best form of government	Kim Jong Un inherited his power from Kim Jong Il and Kim Il Sung so he has a legitimate claim to a familial dynasty . Kim Jong Un would be much more inclined to agree to this than any other solutions that would unify the Korean peninsula . He retains a luxurious life style and standard of living and his rule and kingdom are expanded further .	5
3	People should not be able to purchase soda with SNAP benefits	For example , other foods that have excessive preservatives or known carcinogens .	3

Table 5.6: Examples of **Global Consistency** (Stance). Example 1 would score a 1 for internal consistency, as it switches positions every sentence, but a 3 for stance. We instructed annotators "You should also consider that the stance may be explicitly stated," e.g. "I don't think X," but it may be implicit as well. If it is implicit, there should be a reasonable number of implicit inference steps. Example 2 may be implicitly arguing that all monarchies are bad by providing an example of one, so it could receive a high score (5). On the other hand, Example 3 requires some additional reasoning steps to get to the idea that "SNAP covers unhealthy foods, so it's fine that it covers soda."

Example	Main Claim	Counter-Argument	Score
1	The United States should not cancel student loan debt	You 're a metaphorical gun pointed to your head by many . It 's a " either you 'll flip burgers all your life for 7$/hour or you accept to be indebted a scandalous amount of money to use " . As such , cancelling student debt is deemed as acceptable because society as a whole created a situation where you were forced to accept a scandalous debt .	4/5
2	Welfare = theft	Capital generates money for its owner just by being an owner , and not through any productive labor .	5
3	If you are pro - choice then you should also believe in the right to assisted suicide	I do n't think the right to assisted suicide is a right to life . I think it 's a right for people to end their life early if they wish to end it early . It 's not a right that should be given to everyone . It should only be granted to those who have the right .	1
4	People should not be able to purchase soda with SNAP benefits	The purpose of SNAP is not to supplement nutrition , it 's to help people get the nutrients they need to survive . If you do n't drink soda , you wo n't be able to get the nutrition you need to live .	2

Table 5.7: Examples of argument **Quality**. Example 1 may be considered a fairly high-quality argument, as it states the reasoning for the counter-argument by using a metaphor. When considering whether an argument makes a good point, it is important to consider the context. In other words, the argument may directly or indirectly (through implicit reasoning) refute the main claim of the argument. For example, given the main claim The United States should not cancel student loan debt, the counter-argument *"How do you feel about declaring bankruptcy?"* would receive a high score for quality, as bankruptcy cannot be declared for student loan debt. Example 2 is similar, as the argument is claiming that the owner of a business is just like a welfare recipient, as they are receiving benefits for no work. Finally, the argument may not have a consistent chain of reasoning. Example 3 concludes with an inconsistency and fails to provide any reasoning. Example 4 starts with a reasonable statement, but the last sentence is not a factual statement.

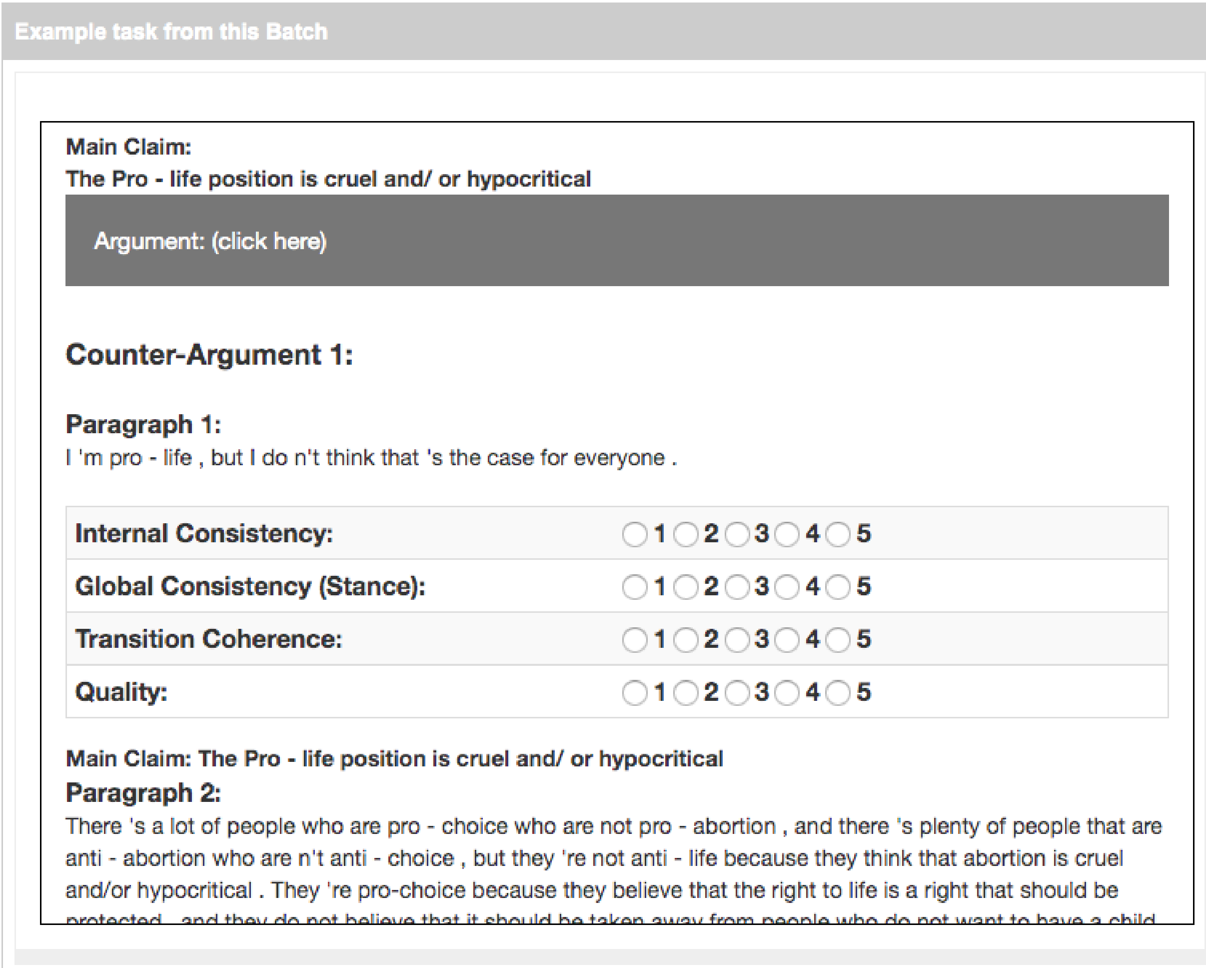


Figure 5.13: An example of the interface presented to annotators.

paragraphs for Baseline-BART.

The results of the evaluation are presented in Table 5.8. Overall, the annotators scored the arguments of Our Hybrid System much higher in terms of Quality. Baseline-BART tends to generate simple, often contradictory chains of reasoning. A system with a retrieval-based component is better able to provide complex arguments.

Our Hybrid System also performs much better at Internal Consistency, which is not surprising, as the retrieved paragraphs were mostly kept intact during the generation component. On the other hand, Baseline-BART performs significantly worse than Our Hybrid System, suggesting that the model has a limited understanding of semantics when it comes to contrast. We suspect that Our Hybrid System receives less than a perfect score due to the BART-based fusion component.

Model	Internal	Global (Stance)	Transition	Quality
Our Hybrid System	4.51	2.99	3.71	3.02
Baseline-BART	3.78	2.8	3.4	2.41

Table 5.8: The results of our human evaluation of Our Hybrid System (the end-to-end approach) and Baseline-BART.

In addition, Our Hybrid System obtains comparable performance to BART-Baseline in terms of Transition Coherence, which is somewhat surprising as Baseline-BART was trained end-to-end. This performance suggests that our approach of replacing and inserting argumentative shells is reasonable.

However, stance detection is a difficult problem for both models. Baseline-BART has mediocre scores for stance in part because it generates many different stances (hence the low score for internal consistency). On the other hand, although Our Hybrid System obtains scores for stance similar to other work in argument generation [Hua et al., 2019a, Hua and Wang, 2019], there is significant room for improvement. One reason why may be that our assumption that every response can be used as a counter-argument is not always applicable, especially when the response is a clarifying question. Furthermore, while the stance may be correct at the *post-level*, the *paragraph-level* provides additional difficulty, as individual paragraphs may be non-argumentative or even briefly take the opposing view as a prebuttal technique. Our methods for argumentative component detection help to identify arguments, but the resulting paragraphs still contain some errors. Additionally, improvements to argumentative relation detection (along with the type of relation – support or attack) would improve the fine-grained detection of arguments.

Ablation of System Components To further understand where the performance differs from the BART-Baseline, we conduct additional experiments where we remove or replace certain components in the pipeline. First, we examine only the cases where the selected paragraph is actually a counter-argument according to the gold labels (170 cases, or 50% of the time). The annotators score for Global Consistency (Stance) improves to 3.62, showing that the expert annotators tend to agree with the labels provided by the crowd-workers and that improvements in stance detection will lead to improvements in counter-argument generation. The scores in this case are not higher for the same reasons as the stance scores in the full evaluation: 1) the persuasive argument extraction component may select paragraphs with the wrong stance and 2) the argument component identification component (predicting claims and premises) may allow for paragraphs that are not arguments due to incorrect predictions. Future work could examine joint models of stance and persuasiveness due to this limitation.

Second, we examine only the cases where the paragraph is a counter-argument *and* is persuasive according to whether the response received a delta. When we examine only the cases that received a delta but were not counter-arguments, we find that annotators assigned these examples lower scores for Quality (2.8) because they also had low scores for Global Consistency/Stance (2.6). When both conditions apply, the scores for Quality increase to 3.2. While the labels for persuasion were provided by different posters in response to a different original post, these labels should be a proxy for overall quality.

Finally, we conduct an additional evaluation using the same annotators on a subset of 25 full arguments[20] from our full human evaluation. We evaluate two alternatives to the persuasive argument extraction component. For the first, we evaluate the use of a **paragraph embedding** approach, where we represent each paragraph and provided OP as a weighted average of GloVe embeddings, and select the T paragraphs with the maximum cosine similarity to the OP. To order the selected paragraphs, we compute the sum of cosine similarities between paragraphs for all possible $T!$ sequences, which is feasible as $T < 5$ in practice. For the second alternative, we compare to a **retrieval-based** baseline, where instead of extracting the most persuasive paragraphs from all candidate counter-arguments, we use our persuasive argument model to rank entire counter-arguments and select the first three paragraphs from the highest scoring one. This baseline is analogous to the LEAD-3 baseline in extractive summarization, which has been shown to be surprisingly difficult to outperform [Kryscinski et al., 2019].

The results of this ablation study, compared to the results of Our Hybrid System and Baseline-

[20] 62 paragraphs for Our Hybrid System and 54 for Baseline-BART

Model	Internal	Global (Stance)	Transition	Quality
Our Hybrid System	4.85	3.37	3.79	3.15
Baseline-BART	3.7	2.8	3.26	1.69
Paragraph Embeddings	4.82	2.85	3.15	2.49
Retrieval-based	4.94	2.27	3.41	1.87

Table 5.9: The results of our ablation studies for the persuasive argument extraction component. We compare the human evaluation of Our Hybrid System (the end-to-end approach) and Baseline-BART to a system that uses average word embeddings (Paragraph Embeddings) and one that applies the persuasion model to an entire post (Retrieval-based).

BART on the same examples, are presented in Table 5.9. Notably, both the embedding approach and the retrieval-based approach result in lower quality, suggesting that our persuasion model is effective at both identifying persuasive paragraphs and that our approach of representing the "best" arguments against a topic is better than representing only those viewpoints from the same author. Furthermore, as we select the first three paragraphs in the retrieval-based approach regardless of their argumentative content, this may result in paragraphs with a high number of non-argumentative sentences and unsupported claims, leading to lower scores for both stance and quality. This issue is mitigated somewhat by the paragraph embeddings, which select paragraphs according to content and are likely to remove some non-argumentative paragraphs but do not explicitly consider argumentation. Another interesting result is the lower scores for transition coherence for both the retrieval-based and paragraph embedding approaches. In the case of paragraph embeddings, the fact that the persuasion model may be a better model of coherence is one possible explanation. However, we would expect the retrieval-based model to have natural transitions since we use a full argument exactly as it was written by the author. Further examination of the scores reveals that the first paragraph tended to score lower than the others (a score of 2.45 compared to 4.0). As we are retrieving counter-arguments that may not be directly compatible with the original post, the fusion component is essential for providing a good introductory transition.

Qualitative Analysis To better understand the performance of the model, we present output from both models along with the title claim and part of the original post.

In the example in Table 5.10, Our Hybrid System provides three solid arguments from different sources, claiming that 1) student loans can be discharged if a school was discredited, so a credited school should have the same options 2) student loan debt is a systemic problem limiting economic output and 3) most loan recipients are young and should not be held responsible. The fusion model also learns to enumerate the arguments. One issue with this output, however, is that the model generates *"i 'm not sure if"* to connect to *"i 'm aware,"* which is not coherent. A better approach would allow for editing of the entire paragraph.

Compared to Our Hybrid System, Baseline-BART makes some reasonable points but is unable to maintain a consistent stance for an entire paragraph. In the example in Table 5.10, Baseline-BART claims that *"We should make it so that people ca n't afford to go to college,"* but then later states *" I 've been able to pay it off over the course of my life."* Additionally, Baseline-BART tends to be repetitive, re-stating points such as *" It 's also going to be a lot harder to pay off the debt than it would be to pay down the debt , and it 'll be harder to find a way to pay that debt off."*

In Table 5.11, the retrieved arguments are reasonable counter-arguments against paperclips - *"pages can come loose," "If you drop something bound by only a paperclip , you may well have dozens of pages scattered all over the floor,"* and *"they are not tamper proof"* in each respective paragraph. However, the BART-based fusion component inserts claims arguing against staplers - e.g. *"Staplers are not tamper proof"*. On this example, Baseline-BART seems to struggle to make a coherent point, claiming that *"StaplERSare more expensive and require more paperclips"* and *"the paperclips are more likely to get stuck to the paperclip."*

The limitations of the retrieval approach are also evident in Table 5.12. Our Hybrid System actually provides evidence in favor of the original claim, rather than a counter-argument, stating that *"America has an obesity problem,"* which could be addressed by restricting the purchasing of soda. If we use gold stance labels instead of our stance detection model, the model provides a solid argument, making an analogy to the consumption of information and implying that if paternalism in one area is bad then it is also bad for restricting consumption.

Overall, in two of the three examples, Baseline-BART starts the argument with *"I'm not sure what you're talking about here."* Baseline-BART has a tendency to generate generic phrases at the start of the argument, which are not helpful as meaningful counter-arguments but may serve a rhetorical purpose. Using BART as a fusion component in a hybrid system allows us to generate generic phrases while also providing content via the retrieval mechanism.

Baseline-BART also tends to generate false statements. In Table 5.11, Baseline-BART claims that staples are more expensive than paperclips, where a quick fact-check reveals that paperclips are more expensive. Baseline-BART also provides the argument *"If you do n't drink soda , you wo n't be able to get the nutrition you need to live."*

5.1.5 Conclusions, Limitations, and Future Work

We developed a hybrid approach to providing counter-arguments, using a pipeline of three components: stance detection, persuasive argument extraction, and fusion. Our approach outperforms a state-of-the-art model for text-to-text generation on a number of different measures.

However, a number of areas remain for improvement. First, stance detection is a difficult problem. Our ensemble approach works reasonably, but the model would benefit significantly from improvements in stance detection. One possibility would be to train a graph convolutional network [Bruna et al., 2014] on the directed acyclic graph of the discussion threads and title claim relations. This approach may allow us to learn the constraints of the graph (e.g. that responses should all have the same relation), rather than training each edge independently.

Additionally, improvements to our argumentative relation prediction model from Section 4.2.2 would also lead to improvements on argument extraction. Our model considers relations at a coarse-grained level – between posts – but ideally we would be able to identify the relation between *every* claim and premise. Improvements in this area also open up the possibility of using other argumentative units rather than paragraphs.

Next, there is significant room for improvement on persuasion prediction. While our experiments in Section 4.3.3 showed that persuasive influence is a difficult problem for humans, given larger datasets and better models we may be able to obtain improved performance. Other aspects of persuasion one may consider include framing and personality traits.

Title Claim	The United States should not cancel student loan debt
Our Hybrid System	**i 'm not sure if** ~~but and as far as~~ i 'm aware , most private educational debt is dischargeable if the debt incurred in attending , essentially , not a " real " school . if the educational institution has gotten over that hurdle and is licensed just the same as a " non - profit " school , why should the debt be treated any different if the school is for profit or not ? (and on a very base level , there are no such things as non - profit schools anyways , since their existence is founded on a belief that they provide positive returns to the broader society in the form of an enlarged tax base) **[PAR] 2 .)** Have you heard the saying , " when you owe the bank $ 100 , that 's your problem , when you owe the bank $ 1,000,000 , that 's the bank 's problem ? " That 's what we 're talking about here . It 's less a prescription for erasing an individual 's debt or responsibility and more looking at the global ramifications to having an entire generation under the thumb of financial agencies with an inability to contribute in a diffuse way to economy stimulation . **[PAR] 3 . For many people ,** ~~In addition, for the vast majority of kids~~ this is their first loan . Meaning THEY HAVE NEVER HAD A LOAN BEFORE ! If you 've never had a loan , have no idea what is and is n't a good interest rate , continue to hear from admins , loan advisors and the federal government that taking it makes sense how do you expect that people stop taking " bad " loans ?
Baseline-BART	I 'm not sure what you 're talking about here . **[PAR]** You 're right that we should n't cancel student loan debt , but that 's not the only way to do it . We should make it so that people ca n't afford to go to college . I 'll give you an example . I went to college for two years . I graduated with a degree in political science . I had a lot of debt when I graduated , but I 've been able to pay it off over the course of my life . I did n't have to worry about paying it back . I was able to get a job that paid me a living wage . I could afford to live on my own for a year or two and then I could go back to school . I do n't know if I 'd be able to afford to do that if I had to pay back the debt . **[PAR]** The problem with this is that it 's going to cost a lot more than just the cost of the education . It 's also going to be a lot harder to pay off the debt than it would be to pay down the debt , and it 'll be harder to find a way to pay that debt off . I think the best way to fix this is to make it easier for people to get out of debt . If you can pay off your debt , you can get a good job . If not , then you 'll have to find another job that pays you a livable wage .

Table 5.10: An example of output from Our Hybrid System and Baseline-BART for the given Title Claim.

Title Claim	Paperclips are better than staplers
Our Hybrid System	**Staplers make it easier for the paper to come off , but** ~~If~~ pages can come loose , people will lose the position of the active slide , it can create a mess and potentially interruptions . **[PAR] Staplers make it easier for the paper to come off ,** Staples are perhaps mildly less convenient for the producer , but I 'd argue far favorable to the consumer in most cases . If you drop something bound by only a paper clip , you may well have dozens of pages scattered all over the floor , getting out of order and just a pain to collect , re - order , and rebind . The permanence of a staple is a benefit in that scenario . **[PAR] Staplers are not tamper proof .** I.e. they are not tamper proof
Baseline-BART	Staplers are better than paperclips because they are easier to use . **[PAR]** Paperclips are better because they do n't require staples . Staplers require staples , and paperclips require a lot of paperclips . They are also easier to clean and easier to get out of the way , and they are more reusable . StaplERS are more expensive and require more paperclips , and are more difficult to clean . Paperclips can be reusable , but they are not as easy to clean as a stapler , and the paperclips are more likely to get stuck to the paperclip . They can also get stuck on the paper , which can make it difficult to get them out of your pocket when you need to use them . They 're also more expensive .

Table 5.11: An example of output from Our Hybrid System and Baseline-BART for the given Title Claim.

Title Claim	People should not be able to purchase soda with SNAP benefits
Our Hybrid System	**I think you 're right .** ~~In my opinion, the only reason~~ America has an obesity problem (more significant then most other developed countries) is because of healthy food being WAY too expensive . Sugary , fatty food is so much cheaper then healthier meals . **[PAR] I think** I would be happy if this was made a reality . It would really deter me from buying frivolous unhealthy foods . Should my diet call for the occasional unhealthy food , I 'm happy to pay a little more for it . **[PAR] I think you 're right .** I realize it would result in some big players maybe going out of business but I think it 's more likely they just change their business model and they would adapt and overcome .
Baseline-BART	I 'm not sure what you 're talking about here . **[PAR]** You 're right that soda has no nutritional value , but that 's not the point . The purpose of SNAP is not to supplement nutrition , it 's to help people get the nutrients they need to survive . If you do n't drink soda , you wo n't be able to get the nutrition you need to live . You 'll just be wasting your money on something you ca n't get from a grocery store . If I were you , I would n't have the money to buy a candy bar or a soda . I 'd be spending my money on food . I would be wasting my time and energy on something that I can not get from the grocery store , and I 'll be wasting it on something I can n't afford to buy from the store . It 's a waste of money .
Gold Stance	**I mean ,** I guess we should also limit people's intake of information because all classes of people tend to misuse the information systems that are available, consuming (and regurgitating) mass amounts of junk information, which ultimately has a negative effect on the entire society (and world). I get to decide what information you're allowed to access. Progressive reform. **[PAR] I mean ,** What about this reform is "progressive", exactly? Like, Iḿ sidelining your argument about the policy change and just directly asking how you mesh limiting food stamp benefits and a mindset that benefits are some sort of privilege that people squander with "liberal on economic issues" and progressive ideas. It's like... exactly a Conservative argument. **[PAR] 2.** ~~Another consideration is the mindset of those with generational poverty.~~ Ruby Payne's A Framework for Understanding Poverty is a great read to kind of understand the motivation behind some of the choices people in poverty make. When you're in poverty, you're more about the right here and right now and money is to be used and spent. It's pretty interesting.

Table 5.12: An example of output from Our Hybrid System and Baseline-BART for the given Title Claim. We also present a comparison to Our Hybrid System when given Gold Stance labels.

Furthermore, rather than re-writing only the argumentative shells between paragraphs, future improvements could re-write the entire paragraph. Previous work showed that matching style, i.e. entrainment, between the non-content words of a response and an original post is more likely to lead to a convincing argument [Tan et al., 2016]. Similarly, *generating* a response that matches style is one direction for future work. Other possibilities could include the generation of *concessions* [Musi, 2018], where a component of the system learns to agree with an aspect of a post, but then generate an alternative view. Another aspect of re-writing could involve the use of *pathos*, as discussed in Section 4.1.2. As our analysis showed that logos and pathos are often used jointly in convincing arguments, one possibility is that the use of pathos mostly consists of lexical choice. If this is true, modifying arguments by replacing words with emotional *connotations* (e.g. by using fine-grained lexicons [Rashkin et al., 2016]) may have the effect of making an argument more persuasive. Furthermore, emotional effectiveness may be topic-dependent and so these words would have different connotations in different contexts.

Finally, there is potential to use BART in alternative ways to our system. Our fusion approach assumed that all non-content phrases can be generated by a single model, when we may want to separately model discourse, speech acts, or hedging. Next, as our hybrid system is agnostic to the source of arguments, future work could also experiment with using BART as one of these sources. Our qualitative analysis showed that BART is sometimes capable of generating quality points, but often becomes incoherent for longer contexts. An extractive approach could identify the quality arguments from the output of a BART model, e.g. by using an automated fact checking system [Hidey and Diab, 2018] or verifying that the premises support the claims using an approach such as the one described in Section 4.2.2.

5.2 Contrastive Claim Generation

In Section 5.1, we discussed how to retrieve counter-arguments given a claim using models for stance detection. For example, for the claim *"Capitalism is the best economic system,"* we might obtain counter-arguments that provide evidence in support of socialism. In this section, we instead describe a method to modify a claim by making semantic edits that result in a claim with contrastive meaning, e.g. this method allows us to directly generate *"Socialism is the best economic system"* by making a meaningful replacement. In Section 5.1 we discussed how we made replacements to insert the appropriate discourse connectives between arguments, an operation that is largely functional and does not modify the stance. Here we modify the content itself by learning a method for contrastive lexical substitution.

Contrast is a key component of argumentation. In the model of Toulmin [1958], used for our analysis of arguments in Sections 4.1 and 4.2, the center of the argument is the claim, a statement that is in dispute [Govier, 2010]. As part of the Toulmin model, a rebuttal is a necessary component in a dialogue. In Section 4.2, we discussed that rebuttal is one method for directly challenging the truth of a claim. In order to rebut a claim, we then need a method to generate a counter-claim with contrastive meaning to the claim of interest.

As direct rebuttal to a claim is one possible argumentative strategy, counter-claim generation is one possible step in an argument generation process. While alternatives exist, such as our approach discussed in Section 5.1, which directly retrieves counter-arguments and thus implicitly states counter-claims, we may desire a method that explicitly states the opposing stance for rhetorical purposes. In particular, Apothéloz et al. [1993] find that explicit counter-claims can help support the conclusion of the counter-argument.

Given an argument, a system that generates counter-arguments may need to 1) identify the claims to refute, 2) generate a new claim with a different view, and 3) find supporting evidence for the new claim. We focus on this second task of counter-claim generation. This approach is complementary to that in Section 5.1, where we retrieved counter-arguments for a claim. Instead, we could generate a counter-claim and retrieve arguments in support of the claim, using the same approach.

We build on previous work in counter-claim generation [Bilu et al., 2015], which focused on explicit negation to provide opposing claims. While negation plays an important role in argumentation [Apothéloz et al., 1993], researchers found that explicit negation may result in incoherent responses [Bilu et al., 2015]. Furthermore, recent empirical studies have shown that arguments that provide new content [Wachsmuth et al., 2018b] tend to be more effective. While new concepts can be introduced in other ways by finding semantically relevant content, we may find it desirable to explicitly model contrast in order to control the output of the model as part of a rhetorical strategy, e.g. concessions [Musi, 2018]. We thus develop a model that generates a contrastive claim given an input claim.

Bilu et al. [2015], in their work on explicit negation of claims, noted that not every claim has an exact opposite. Consider a claim from Reddit: *"Easy solution, get employers out of the business entirely, **pass universal single payer healthcare**."* This is an example of a policy claim - a view on what should be done [Park et al., 2015, Schiappa and Nordin, 2013] and one of the types of claims discussed in Section 4.1.1 – interpretation. While negation of this claim is a plausible response (e.g. asserting there should be no change by stating *"Do not get employers out of the business, do not pass universal healthcare"*), negation limits the diversity of responses that can lead to a productive dialogue. Instead, consider a response that provides an alternative suggestion: *"Easy solution, get employers out of the business entirely, **deregulate and allow cross-state competition**."* In this claim, the speaker believes in a decreased role for government while single-payer healthcare would result in an increased role. As these views are on different sides of the political spectrum, it is unlikely that a single speaker would utter both claims.

In related work, de Marneffe et al. [2008] define two sentences as contradictory when they are extremely unlikely to be true simultaneously. We thus define a contrastive claim as one that is *likely to be contradictory if made by the speaker of the original claim.* Our goal, then, is to develop a method for generating contrastive claims when explicit negation is not the best option. Generating claims in this way also has the benefit of providing new content that can be used for retrieving or generating supporting evidence.

In order to make progress towards generating contrastive responses, we need large, high-quality datasets that illustrate this phenomenon. We **construct a dataset** of 1,083,520 contrastive comment pairs drawn from Reddit and **build a predictive model** to filter out non-contrastive claims.

Each pair contains very similar, partially aligned text but the responder has significantly modified the original post. We use this dataset to model differences in views and **generate a new claim given an original comment**. The similarity within these pairs allows us to use them as distantly-labeled, high-quality contrastive word and sentence alignments within our model. The word alignments provide semantic information about which words and phrases can be substituted *in context* in a coherent, meaningful way.

Our contributions[21] are as follows:

1. Methods and data for *contrastive claim identification* to mine comment pairs from Reddit, resulting in a large, continuously growing dataset of 1,083,520 distant-labeled examples.

2. A crowd-labeled set of 2,625 comments each paired with 5 new contrastive responses generated by additional annotators.

3. Models for *generating contrastive claims* using neural sequence models and constrained decoding.

In Section 5.2.1 we present background on the task and describe our data collection and processing steps in Section 5.2.2. Next, we present neural models for contrastive claim generation 5.2.3. Finally in Sections 5.2.4 and 5.2.5, we present our experiments and results along with an analysis of the model output.

5.2.1 Background

Contrastive claims may differ in more than just viewpoint; they may also contain stylistic differences and paraphrases, among other aspects. We thus propose to model contrastive claims by controlling for context and maintaining the same text between pairs of contrastive claims except for the contrastive word or phrase. Much of the previous work in contrast and contradiction has examined the relationship between words or sentences. In order to understand when words and

[21] Data and code available at `https://github.com/chridey/fixedthat`

phrases are contrastive in argumentation, we need to examine them *in context*. For example, consider the claim *"Hillary Clinton should be president."*[22] A reasonable contrastive claim might be *"Bernie Sanders should be president."* (rather than the explicit negation *"Hillary Clinton should not be president."*) In this context, Hillary Clinton and Bernie Sanders are contrastive entities as they were both running for president. However, for the claim *"Hillary Clinton was the most accomplished Secretary of State in recent memory."* they would be unrelated. Consider also that we could generate the claim *"Hillary Clinton should be senator."* This contrastive claim is not coherent given the context. Generating a contrastive claim then requires 1) identifying the correct substitution span and 2) generating a response with semantically relevant replacements.

While some contrastive claims are not coherent, there are often multiple plausible responses, similar to tasks such as dialogue generation. For example, *"Donald Trump should be president"* is just as appropriate as *"Bernie Sanders should be president."* We thus treat this as a dialogue generation task where the goal is to generate a plausible response given an input context.

5.2.2 Data

In order to model contrastive claims, we need datasets that reflect this phenomenon.

Collection: We obtain training data by scraping the social media site Reddit for comments containing the acronym *FTFY*.[23] As described in Section 3.1.2, FTFY is a common acronym meaning "fixed that for you."[24] FTFY responses (hereafter FTFY) are used to respond to another comment by editing part of the "parent comment" (hereafter parent). For example, if the parent is *"**Bernie Sanders** for president"*, one possible FTFY is *"**Hillary Clinton** should be president."* To obtain historical Reddit data, we mined comments from the site pushshift.io for December 2008 through October 2017. This results in 2,200,258 pairs from Reddit, where a pair consists of a parent and an FTFY. We find that usage of FTFY began increasing significantly around 2009-2010 until reaching

[22] The *temporal context*, in this case the 2016 United States presidential election, may be relevant for understanding this claim but it was not a focus of this work.

[23] `https://en.wiktionary.org/wiki/FTFY`

[24] `https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/what-do-all-these-acronyms-mean`

Subreddit	Count	Category
askreddit	261771	Discussion
funny	108800	Humor/Parody
pics	92320	Pictures
politics	67282	Politics
wtf	51090	General
leagueoflegends	49962	Gaming
gaming	48940	Gaming
adviceanimals	46039	Memes
worldnews	38260	World News
todayilearned	36088	Discussion

Figure 5.14: Subreddits and Counts

34,057 in October 2017 alone. We show the subreddits that use FTFY most frequently along with their top-level categories[25] in Table 5.14. We find that many of the top occurring subreddits contain topics where we would expect strong opinions (/r/politics, /r/worldnews, and /r/gaming).

Most commonly, FTFY is used for three categories of responses: 1) making a joke, 2) correcting a typo, and 3) expressing disagreement (sometimes sarcastically). We present examples in Table 5.15 along with the subreddit and label. Rows 1 and 3 contain jokes, whereas 2 and 4 contain disagreement. Note that row 4 uses the same modification as row 3, but sarcastically disagrees with the original comment, resulting in a different stance. Finally, row 5 has a correction for the word "piqued."

Classification: To filter the data to only the type of response that we are interested in, we annotated comment pairs for contrastive claims and other types. We use our definition of contrastive claims based on contradiction, where both the parent and FTFY are a claim and they are unlikely to be beliefs held by the same speaker. A joke is a response that does not meaningfully contrast with the parent and commonly takes the form of a pun, rhyme, or oronym. A correction is a response to a typo, which may be a spelling or grammatical error. Any other pair is labeled as "other," including pairs where there is no difference between the parent and FTFY (i.e. the original commenter edited the parent in response to the FTFY), the parent is not a claim, or there is no relationship

[25]Categories obtained from the `https://snoopsnoo.com` API

Parent	FTFY	Subreddit	Label
This **Python library** really **piques** my interest.	This really ***py*ques** my interest.	programming	joke
Screw the **yankees**.	Screw the **red sox**.	baseball	disagreement
And to be fair to Kasich, he's not **Ted Cruz**, so he has that going for him	And to be fair to Kasich, he's not **the Zodiac killer**, so he has that going for him	politics	joke
Ted Cruz for president	**The Zodiac killer** for president	politics	disagreement
i have seen a few rp focused posts here recently and it has peaked my curiosity	**piqued** my curiosity	wow	typo

Figure 5.15: Examples of FTFYs

between the parent and FTFY (which may be due to a bad alignment).

In order to identify contrastive claims, we selected a random subset of the Reddit data from prior to September 2017 and annotated 1993 comments. Annotators were native speakers of English and the Inter-Annotator Agreement on 300 samples using Kripendorff's alpha was 0.72. The distribution of the labels is presented in Table 5.16. Contrast occurs in slightly more than half of the sampled cases (51.4%), with jokes (23.0%) and corrections (21.2%) comprising about one quarter each. We then train a binary[26] classifier to predict contrastive claims, thus enabling better quality data for the generation task.

Label	Number	Percentage
Contrast	1026	51.4%
Joke	460	23.0%
Other	423	21.2%
Correction	83	4.2%

Figure 5.16: Distribution of Labels

To identify the sentence in the parent that the FTFY responds to and derive features for classification, we use an edit distance metric to obtain sentence and word alignments between the parent comment and response; this is needed because the FTFY is often shorter than the parent, where

[26]We combine all non-contrast classes into a single class.

the parent consists of multiple sentences and the FTFY only a single sentence. As the words in the parent and response are mostly in the same order and most FTFYs contain significant overlap with the parent response, it is possible to find alignments by moving a sliding window over the parent. We select alignments with the best score using edit distance, and extend the alignment to the nearest sentence boundary for both parent and FTFY so we have full sentences for modeling. A sample of 100 comments verifies that this approach yields exact word alignments in 75 comments and exact sentence alignments in 93.

Given these pairs of comments, we derive linguistic and structural features for training a binary classifier. For each pair of comments, we compute features for the words in the *entire comment span* and features from the *aligned phrases span* only (as identified by edit distance). From the *aligned phrases* we compute the *character edit distance* and *character Jaccard similarity* (both normalized by the number of characters) to attempt to capture jokes and typos (the similarity should be high if the FTFY is inventing an oronym or correcting a spelling error). From the *entire comment*, we use the *percentage of characters copied* as a low percentage may indicate a poor alignment and the *percentage of non-ASCII characters* as many of the jokes use emojis or upside-down text. We also include the *length* of both the parent and FTFY normalized by the maximum length comment. In addition, we use features from GloVe [Pennington et al., 2014] word embeddings[27] for both the *entire comment* and *aligned phrases*. We include the *percentage of words in the embedding vocabulary* for both spans for both the parent and FTFY. The reason for this feature is to identify infrequent words that may be typos or jokes. We compute the *cosine similarity* of the average word embeddings between the parent and FTFY for both spans. Finally, we use *average word embeddings* for both spans for both parent and FTFY.

As we want to model the generation of new content, not explicit negation, we removed any pairs where the difference was only "stop words." The set of stop words includes all the default stop words in Spacy[28] combined with expletives and special tokens (we replaced all URLs and usernames). We trained a logistic regression classifier and evaluated using 4-fold cross-validation.

[27] We found the 50-dimensional Wikipedia+Gigaword embeddings to be sufficient

[28] `https://spacy.io`

We compare to a *character* overlap baseline where any examples with Jaccard similarity > 0.9 and edit distance < 0.15 were classified as non-contrastive. The goal of this baseline is to illustrate how much of the non-contrastive data involves simple or non-existent substitutions. Results are shown in Table 5.13. Our model obtains an F-score of 80.25 for an 8 point absolute improvement over the baseline.

Model	Precision	Recall	F-score
Majority	51.4	100	67.5
Baseline	67.75	77.19	72.16
LR	74.22	87.60	80.25

Table 5.13: Results of Identifying Contrastive Claims

Selection: We select the final dataset such that the FTFY length is between 2 and 50, the parent length is between 4 and 50, and the ratio of the parent and FTFY lengths is less than 9 to 1. We chose these settings to prevent the model from generating short responses or compressing the output significantly. As we are primarily working with sentences, the mean parent length was 16.3 and FTFY length was 14.3. We also found that these settings did not reduce the F1 score on our labeled data. Given these settings, we use our trained model to classify the remaining data and obtain 1,083,797 Reddit pairs. We set aside 10,307 pairs from October 1-20, 2017 for development and October 21-30 for test (6,773), with the remainder used for training.

The resulting test FTFYs are naturally occurring and so do not suffer from annotation artifacts. At the same time, they are noisy and may not reflect the desired phenomenon. Thus, we also conducted an experiment on Amazon Mechanical Turk[29] (AMT) to obtain additional gold references, which are further required by metrics such as BLEU [Papineni et al., 2002] or METEOR [Banerjee and Lavie, 2005]. We selected 2,625 pairs from the 10 most frequent categories[30] (see Table 5.14). While Table 5.14 shows the top-level categories, many top-level categories are overly broad. The categories form a three-level hierarchy for each subreddit and we use the second-level, e.g. for

[29] We paid annotators per HiT at a rate commensurate to the U.S. federal minimum wage and the study was approved by the Columbia Institutional Review Board (IRB-AAAS1977).

[30] Again using the `https://snoopsnoo.com` API

Verify stances that disagree with a reddit comment

You will be shown a reddit comment and 2 similar comments with different stances.
You will need to select which of 2 choices makes more sense based on the context.

Select the response that makes the most sense compared with the comment "browns fans are going to be insufferable this week ."

- steelers fans are going to be insufferable this week .
- orioles fans are going to be insufferable this week .

Select the response that makes the most sense compared with the comment "the greatest play in super bowl history"

- the greatest play in nfl history
- the greatest play in nba history

Select the response that makes the most sense compared with the comment "beat miami for us now"

- beat california for us now
- beat new england for us now

Figure 5.17: Sample Questions from Qualification Test Given to Mechanical Turk Workers

/r/pokemongo the categories are “Gaming,” “Video Games,” and “Pokémon” so we use “Video Games.”

Before participating, each annotator was required to pass a qualification test - five questions to gauge their knowledge of that topic. Each question was a binary choice and we required them to answer all questions correctly, as we considered this to be the best tradeoff between annotator time and the likelihood of an unqualified annotator randomly guessing their way to success. For the movies category, for one question we provided the claim *“Steven Spielberg is the greatest director of all time.”* We then asked the annotators to select either *“Stanley Kubrick”* or *“Paul McCartney.”*, as only the former would be a valid alternative. A screenshot of the “Football” qualification test is given in Figure 5.17.

Category	Count	Category	Count
Video Games	1062	Basketball	116
Politics	529	Soccer	99
Football	304	Movies	88
Television	194	Hockey	60
World News	130	Baseball	55

Table 5.14: Comments for Mechanical Turk

If they passed this test, the annotators were then given a parent comment from the aforementioned 2,625 test examples and were asked to generate a new comment by changing it to have the opposite meaning. They were explicitly instructed to avoid simple negation using words such as *"not"* or *"can't"*. Annotators were provided with the subreddit of the comment along with the three category levels to provide additional context. We manually validated each generated FTFY before approval to remove obvious spam or trivial negation. We also informed annotators that if their hit was rejected, they could contact us directly to address their mistakes and we worked with several annotators as a result to improve their performance.

Overall, we obtained *five* new FTFYs for each parent, resulting in 13,125 pairs for evaluation.

5.2.3 Methods

Our goal of *generating* contrastive claims can be broken down into two primary tasks: 1) identifying the words in the original comment that should be removed or replaced and 2) generating the appropriate substitutions and any necessary context. Initially, we thus experimented with a modular approach by tagging each word in the parent and then using the model predictions to determine if we should copy, delete, or replace a segment with a new word or phrase. We tried the bi-directional LSTM-CNN-CRF model of Ma and Hovy [2016] and used our edit distance word alignments to obtain labels for copying, deleting, or replacing. However, we found this model performed slightly above random predictions, and with error propagation, the model is unlikely to produce fluent and accurate output. Instead, we hypothesize that a better approach would incorporate end-to-end techniques from machine translation.

Our Model: We use neural sequence-to-sequence encoder-decoder models [Sutskever et al., 2014] with attention for our experiments. The tokens from the parent are passed as input to a bi-directional GRU [Cho et al., 2014] to obtain a sequence of encoder hidden states h_i. Our decoder is also a GRU, which at time t generates a hidden state s_t from the previous hidden state s_{t-1} along with the input. When training, the input x_t is computed from the previous word in the gold training data if we are in "teacher forcing" mode [Williams and Zipser, 1989] and otherwise is the prediction made by the model at the previous time step. When testing, we also use the model predictions. The input word w_t may be augmented by additional features. In the baseline scenario $x_t = e(w_t)$ where e is an embedding. The hidden state s_t is then combined with a context vector h_t^*, which is a weighted combination of the encoder hidden states using an attention mechanism:

$$h_t^* = \sum_i \alpha_t^i h_i$$

To calculate α_i^t, we use the attention of Luong et al. [2015] as this encourages the model to select features in the encoder hidden state that correlate with the decoder hidden state, which we want because our input and output are similar. Attention is then calculated as:

$$\alpha_t^i = \frac{\exp(h_i^T s_t)}{\sum_{s'} \exp(h_{s'}^T s_t)}$$

Finally, we make a prediction of a vocabulary word w by using features from the context and decoder hidden state with a projection matrix W and output vocabulary matrix V:

$$P(w) = \text{softmax}(V \tanh(W[s_t; h_t^*] + b_w) + b_v)$$

The standard encoder-decoder architecture with attention is depicted in Figure 5.18. We further discuss how we enhance this architecture specifically for our task.

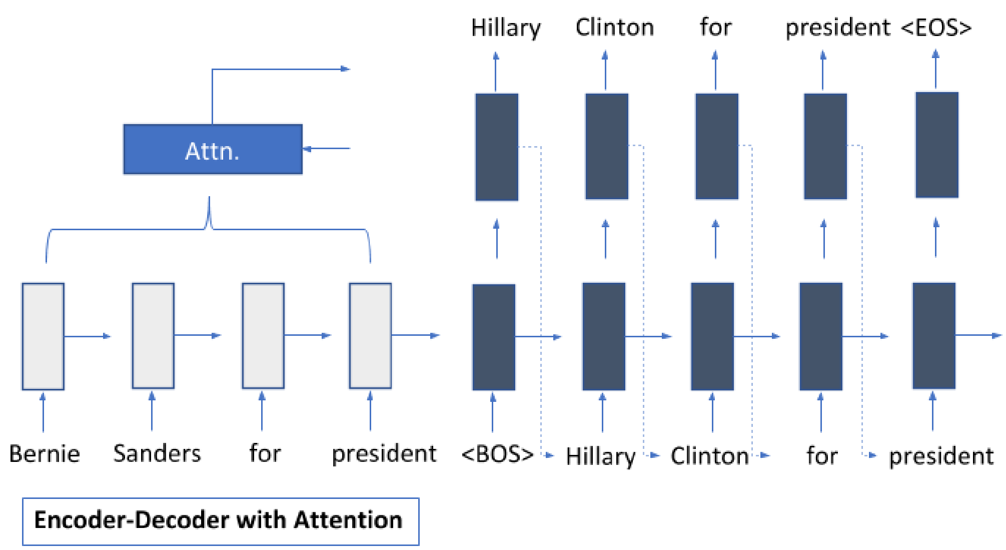


Figure 5.18: The baseline encoder-decoder model with attention. In this example, the source is *"Bernie Sanders for president"* and the target is *"Hillary Clinton for president"*, which is generated auto-regressively conditioned on the attention over the source and the previously generated token.

Decoder Representation We evaluate two representations of the target input: as a sequence of **words** and as a sequence of **edits**. The sequence of words approach is the standard encoder-decoder setup. For the example parent *"Hillary Clinton for president 2020"* and FTFY *"Bernie Sanders for president"* we would use the FTFY without modification. Schmaltz et al. [2017] found success modeling error correction using sequence-to-sequence models by representing the target input as a sequence of edits. We apply a similar approach to our problem, generating a target sequence by following the best path in the matrix created by the edit distance algorithm. The new target sequence is the original parent interleaved with "DELETE-N tokens" that specify how many previous words to delete, followed by the newly generated content. For the same example, *"Hillary Clinton for president 2020"*, the modified target sequence would be *"Hillary Clinton DELETE-2 Bernie Sanders for president 2020 DELETE-1."* We suspect that this approach may help with generating semantic edits as well, as it allows for the replacements to be conditioned on the previously generated sequence of deletions.

Counter Kikuchi et al. [2016] found that by using an embedding for a length variable they were able to control output length via a learned mechanism. In our work, we compute a counter variable, which is initially set to the number of new content words the model should generate. During decoding, the counter is decremented if a word is generated that is not in the source input (I) or in the set of stop words (S).[31] The model uses an embedding $e(c_t)$ for each count, which is parameterized by a count embedding matrix. The input to the decoder state in this scenario is $x_t = e(w_t, c_t)$. At each time step, the count is computed by:

$$c_0 = |O \setminus (S \cup I)| \text{ or desired count}$$

$$c_{t+1} = \begin{cases} c_t - 1, & w_t \notin S \cup I \text{ and } c_t > 0 \\ c_t, & \text{otherwise} \end{cases}$$

[31] the default stop words in Spacy

where O is the set of gold output words in training. We believe this approach will be useful in our scenario, as it allows for the model to plan ahead and determine that it needs to generate words not in the parent comment, rather than just copying the source.

For the parent comment *"Hillary Clinton for president 2020"* and FTFY *"Bernie Sanders for president,"* the decoder input is presented, with the time t in the first row of Table 5.15 and the inputs w_t and c_t in the second and third rows, respectively. At the start of decoding, the model expects to generate two new content words, which in this example it generates immediately and decrements the counter. When the counter reaches 0, it only generates stop or input words.

t	0	1	2	3	4
w_t	-	Bernie	Sanders	for	president
c_t	2	1	0	0	0

Table 5.15: Example of Counter

Unlike the controlled-length scenario, at test time we do not know the number of new content words to generate. However, the count for most FTFYs is between 1 and 5, inclusive, so we can exhaustively search this range during decoding. We experimented with predicting the count but found it to be inaccurate so we leave this for future work.

Subreddit Information As the model often needs to disambiguate polysemous words, additional context can be useful. Consider the parent comment *"this is a strange bug."* In a programming subreddit, a sarcastic FTFY might be *"this is a strange feature."* However, in a Pokémon subreddit, an FTFY might be *"this is a strange dinosaur"* in an argument over whether Armaldo is a bug or a dinosaur. We thus include additional features to be passed to the encoder at each time step, in the form of an embedding g for each the three category levels obtained in Section 5.2.2 (the subreddit and two higher-level categories obtained from Snoopsnoo). These embeddings are concatenated to the input word w_t at each timestep, i.e. $x_t = e(w_t, g_t^1, g_t^2, g_t^3)$.

Copy Prediction We use a negative log likelihood objective function $\mathcal{L}_{NLL} = -\log \sum_{t \in 1:T} P(w_t^*)$, where w_t^* is the gold token at time t, normalized by each batch. We also include an additional loss term that uses the encoder hidden states to make a binary prediction over the input for whether a token will be copied or inserted/deleted. For the same example *"Hillary Clinton for president 2020,"* the target would be $\underline{0\ 0\ 1\ 1\ 0}$. This encourages the model to select features that indicate whether the encoder input will be copied to the output. While we found that performing this task in isolation using a CRF was not helpful, we hypothesize that the model will be able to jointly use the information from the edits along with the generated sequence of words during decoding. We use a 2-layer multi-layer perceptron and a binary cross-entropy loss $\mathcal{L}_{BCE}$. The joint loss is then:

$$\mathcal{L} = \mathcal{L}_{NLL} + \lambda \mathcal{L}_{BCE}$$

This is also a form of multi-task learning.

Constrained Decoding We use beam search for generation, as this method has proven effective for many neural language generation tasks. For the settings of the model that require a counter, we expand the beam by count m so that for a beam size k we calculate $k * m$ states.

We optionally include a constrained decoding mode where we filter the output based on the counter; when $c_t > 1$ the end-of-sentence (EOS) score is set to $-\infty$ and when $c_t = 0$ the score of any word $w \in V \setminus (S \cup I)$ is set to $-\infty$. The counter c_t is decremented at every time step, similar to the counter embedding. In other words, when the counter is zero, we only allow the model to copy or generate stop words. When the counter is positive, we prevent the model from ending the sentence before it generates new content words and decrements the counter. The constrained decoding is possible with any combination of settings, with or without the counter embedding.

The full model is shown in Figure 5.19. In addition to attending over the input at every decoding step, we include additional features passed to the decoder, a copy prediction objective for the encoder, and constrained decoding using beam search.

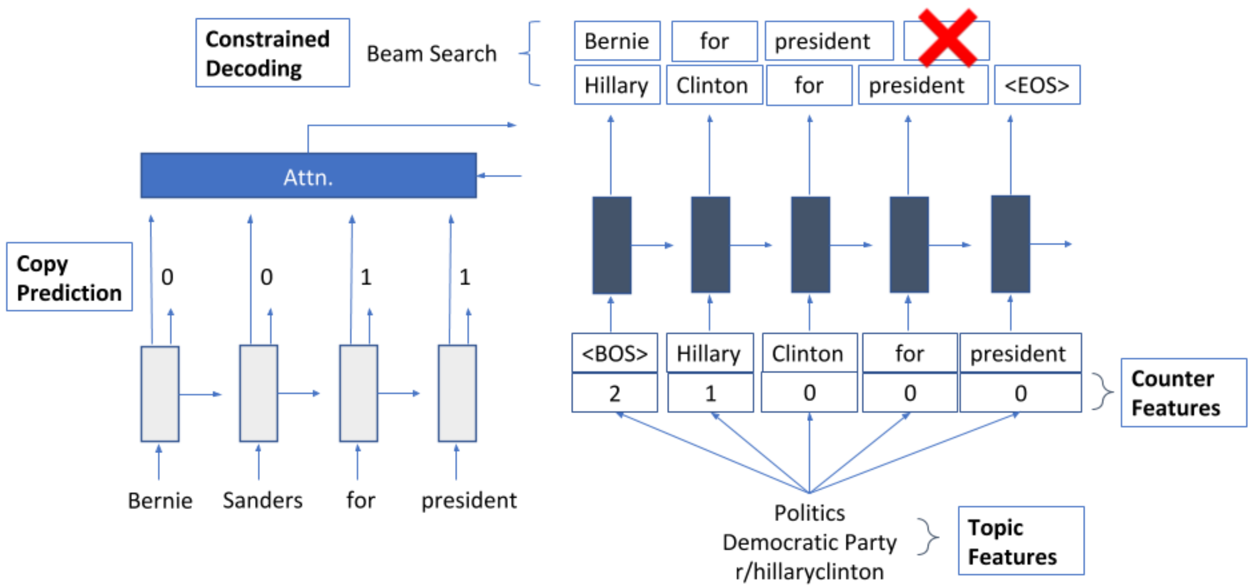


Figure 5.19: Our full model for contrastive claim generation. We encode the source *"Bernie Sanders for president"* and predict whether one of the source tokens will be copied. During decoding, we attend to the source tokens at every timestep, and concatenate the weighted source representations with the previously generated token, the target input, the counter embedding, and the topic features. Finally, we constrain the generated output during beam search so that novel content words are generated.

5.2.4 Experiments and Results

Hyper-parameters and Optimization We used Pytorch [Paszke et al., 2017] for all experiments. We used 300-dimensional vectors for the word embedding and GRU layers. The count embedding dimension was set to 5 with $m = 5$ and $k = 10$ for decoding. The category embedding dimensions were set to 5, 10, and 25 for each of the non-subreddit categories. We also set $\lambda = 1$ for multi-task learning. We used the Adam optimizer [Kingma and Ba, 2015] with settings of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a learning rate of 10^{-3} decaying by $\gamma = 0.1$ every epoch. We used dropout [Srivastava et al., 2014] on the embeddings with a probability of 0.2 and teacher forcing with 0.5. We used a batch size of 100 with 10 epochs, selecting the best model on the development set based on perplexity. We set the minimum frequency of a word in the vocabulary to 4.

For training, development, and testing we use the data described in Section 5.2.2. The test reference data consists of the Reddit FTFYs and the FTFYs generated from AMT. We evaluate our models using automated metrics and human judgments.

Metrics Automated metrics should reflect our joint goals of 1) copying necessary context and 2) making appropriate substitutions. To address point 1, we use BLEU-4 as a measure of similarity between the gold FTFY and the model output. As the FTFY may contain significant overlap with the parent, BLEU indicates how well the model copies the appropriate context. As BLEU reflects mostly span selection rather than the insertion of new content, we need alternative metrics to address point 2. However, addressing point 2 is more difficult due to the variety of possible substitutions, including named entities. For example, if the parent comment is *"jaguars for the win!"* and the gold FTFY is *"chiefs for the win!"* but the model produces *"cowboys for the win!"* (or any of 29 other NFL teams), most metrics would judge this response incorrectly even though it would be an acceptable response. Thus we present results using both automated metrics and human evaluation. As an approximation to address point 2, we attempt to measure when the model is making changes rather than just copying the input. To this end, we present two additional metrics - novelty, a measure of whether novel content (non-stop word) tokens are generated relative to the parent comment, and partial match, a measure of whether the novel tokens in the gold FTFY match any of the novel tokens in the generated FTFY. To provide a reference point, we find that the partial match between two different gold FTFYs (Reddit and AMT) was 11.4% and BLEU was 47.28, which shows the difficulty of automatic evaluation. The scores are lower than expected because the Reddit FTFYs are noisy due to the data collection process described in Section 5.2.2. This also justifies obtaining the AMT FTFYs.

Results are presented in Table 5.16. The baseline is a sequence-to-sequence model with attention. For other components, the counter embedding is referred to as "COUNT," the category/subreddit embeddings as "SUB," the sequence of edits as "EDIT," and the multi-task copy loss as "COPY." The models in the top half of the table use constrained decoding and those in the bottom half are unconstrained, to show the learning capabilities of the models. For each model we compute statistical significance with bootstrap resampling [Koehn, 2004] for the constrained or unconstrained baseline as appropriate and we find the COUNT and EDIT models to be significantly better for constrained and unconstrained decoding, respectively ($p < 0.005$).

			Reddit		AMT	
	Model	Novelty	BLEU-4	% Match	BLEU-4	% Match
Constrained	Baseline	79.88	18.81	4.67	40.14	10.06
	COUNT	89.69	22.61	4.72	47.55	12.55
	COUNT + SUB + COPY	**90.45**	**23.13**	**4.83**	**50.05**	**14.92**
	EDIT	64.64	16.12	3.37	35.48	7.33
	EDIT + COUNT + SUB + COPY	82.96	19.37	4.23	42.69	11.62
Unconstrained	Baseline	3.34	7.31	0.73	25.83	0.68
	COUNT	16.19	8.51	1.95	27.68	2.36
	COUNT + SUB + COPY	16.26	9.62	1.93	31.23	3.81
	EDIT	7.97	**35.41**	1.57	**74.24**	1.56
	EDIT + COUNT + SUB + COPY	**39.99**	32.59	**3.25**	67.56	**6.25**

Table 5.16: Automatic Evaluation

Under constrained decoding, we see that the "COUNT + SUB + COPY" model performs the best in all metrics, although most of the performance can be attributed to the count embedding. When we allow the model to determine its own output, we find that "EDIT + COUNT" performs

Model	Fluency	Coherence	Contrast
Reddit	4.34	4.26	3.01
Baseline (Unconstrained)	3.49	3.19	1.94
Baseline (Constrained)	3.46	3.32	2.53
COUNT + SUB + COPY (Constrained)	**3.52**	**3.46**	**2.87**

Table 5.17: Human Evaluation

the best. In particular, this model does well at understanding which part of the context to select, and even does better than other unconstrained models at selecting appropriate substitutions. However, when we combine this model with constrained decoding, the improvement is smaller than for the other settings. We suspect that because the EDIT model often needs to generate a DELETE-N token before a new response, these longer-term dependencies are hard to capture with constrained decoding but easier if included in training.

We also conducted a human evaluation of the model output on the same subset of 2,625 examples described in Section 5.2.2. We performed an additional experiment on AMT where we asked annotators to rate responses on fluency, coherence, and contrast. **Fluency** is a measure of the quality of the grammar and syntax and the likelihood that a native English speaker would utter that statement. **Coherence** is a measure of whether the response makes sense, is semantically meaningful, and would be usable as a response to a claim. **Contrast** is a measure of how much the response contradicts the original comment. We specified that if the response is different but does not provide a contrasting view it should receive a low rating. Previous work [Bilu et al., 2015] used fluency, clarity/usability (which we combine into coherence), and opposition (where we use contrast).

We used a Likert scale where 5 is strongly agree and 1 is strongly disagree. We used the same data and qualification test from Section 5.2.2 for each category and used three annotators per example. We asked annotators to judge 4 different pairs: 3 model outputs and the gold Reddit[32] FTFYs for comparison. We include the baseline, the baseline with constrained decoding, and the best constrained model ("COUNT + SUB + COPY") according to BLEU and partial match.

[32]We did not evaluate the AMT FTFYs as these were generated by the same pool of annotators.

Parent	Model
ah yes the wonders of **the free market**	ah yes the wonders of **government intervention**
i know that this is an unofficial mod , but **xp** is the best os for this machine	**linux** is the best os for this machine
that 's why it 's important to get all your propaganda from **infowars and brietbart**	propaganda from **fox news outlets**

Table 5.18: Model Output

We verified that the annotators understood how to rate contrast by examining the distribution of responses: the annotators selected option 3 (neither) 15% of the time and preferred to select either extreme, 5 (21%) or 1 (27%). Results are presented in Table 5.17, showing a clear preference for the best model. Note the degradation in fluency for the constrained baseline, as the model is prevented from generating the EOS token and may repeat tokens up to the maximum length. Also note that the contrast is low for the gold Reddit FTFYs, indicating some noise and further justifying the need for AMT FTFYs.

5.2.5 Qualitative Analysis

We provide three examples of the model output in Table 5.18 with the first and third from the News and Politics category, demonstrating how the model handles different types of input. In the first example, the contrast is between allowing markets to regulate themselves versus an increased role of government. In the second example, the contradiction is due to the choice of operating system. In the third (invalid) example, the model responds to a sarcastic claim with another right-wing news organization; this response is not a contradiction since it is plausible the original speaker would also utter this statement.

Error Analysis We conduct an error analysis by selecting model responses from the test set for parent comments where we also had AMT responses. We selected 100 responses where the model did not partially match any of the 6 gold responses and we found 6 main types of errors (some examples given in Table 5.20). One error the model makes is when it identifies an **incorrect**

substitution span while the human responses all selected a different span to replace. We noticed that this occurred 5 times and may require world knowledge to understand which tokens to select. For example, in the first row of Table 5.20, the model changes the target of the claim from *"hillary clinton"* to *"donald trump."* The gold responses all replace the word *"robots,"* suggesting that a better response would have been to change the content after the conditional connective *"if."* Another type of error is when the responses are not coherent with the parent and the language model instead determines the token selection based on the **most recent context** (11 cases). For the claim in the second row, given the most recent output *"still does n't have a"* the model seems to condition more on those four tokens than the parent. We also found examples where the model chose poorly due to unfiltered jokes or **errors in the training data** (12 in total). In 15 cases, the model was not able to find an appropriate response due to the constrained decoding the model and **repeated** a word until the maximum length or appended an incoherent phrase. For the most common error (19 examples), the model made a substitution that **was not contrasting** as in row 3 of Table 5.18 and row 3 of Table 5.20, where the model replaces the word *"iconic"* with a similar word - *"awesome."* Finally, we found 38 of the samples were **valid** responses, but did not match the gold, indicating the difficulty of automatic evaluation. For example, in response to the claim in row 4, the model replaces *"Nintendo"* with *"Gamestop"* (both video game companies) while the gold FTFYs had other video game companies. Similarly, the model replaces *"annoying"* with *"awesome,"* when *"hilarious"* is another acceptable substitution according to the annotators.

5.2.6 Conclusions, Limitations, and Future Work

We presented a new source of over 1 million contrastive claim pairs that can be mined from social media sites such as Reddit. We provided an analysis and models to filter noisy training data from 49% down to 25%. We created neural models for generating contrastive claims and obtained significant improvement in automated metrics and human evaluations for Reddit and AMT test data.

Future work could incorporate this model into an argumentative dialogue system such as that

Type	Parent	FTFY(s)	Model
Span	hillary clinton could have been president if not for **robots**	hillary clinton could have been president if not for [**humans / votes / trump 's tweets**]	**donald trump** could have been president if not for robots
Context	bb-8 gets a girlfriend and poe still does n't have a **boyf** :')	bb-8 gets a girlfriend and poe still does n't have a **girlf** :')	... and poe still does n't have a **clue what i 'm talking about .**
Unchanged	**iconic** ' star trek ' movie memorabilia	**random minor** star trek movie memorabilia	**awesome** ' star trek ' movie memorabilia
Good	**nintendo** is the only company that puts customers over profits	**rockstar** is the only company that puts customers over profits	**gamestop** is the only company that puts customers over profits
Good	it was **annoying** when you drive extremely fast	it was **hilarious** when you drive extremely fast	it was **awesome** when you drive extremely fast

Figure 5.20: Error Analysis

in Section 5.1. In addition to generating claims with a contrasting view, we can also retrieve supporting evidence for the newly-generated claims. Instead of using the title claim as is from Change My View, we could sample a number of contrastive claims and find candidate threads using Google search. Alternatively, we could fact-check generated contrastive claims as an additional step in the counter-argument generation process [Hidey and Diab, 2018].

Our model could be used to generate artificial data to enhance classification performance on tasks such as claim detection (Section 4.2.2) and stance prediction (Section 5.1.2). We also hope that this dataset proves useful for related tasks such as textual entailment [Bowman et al., 2015], providing examples of contradiction, and argument comprehension [Habernal et al., 2018a], providing counter-examples of arguments, or even unrelated tasks like humor or error correction.

Our approach is also limited by the noise in the dataset. Future work could improve the dataset by improving our models for contrastive pair prediction. Another area of improvement is substitution span selection, which is currently handled implicitly. One desired property of this model is controlling which part of the sentence to replace. This would require less noisy labels than the distant-labeled alignments we obtained automatically. Manual labeling could also lead to improvement on the generation task by identifying the types of claims we encounter. For example, we may want to change the target of the claims in some claims but in others change the polarity. In others, we may want to change the content after a discourse connective in a complex clause.

Another way to improve the model is by introducing controllable generation. One aspect of controllability is intention; our model produces contrastive claims without understanding the view of the original claim. Category embeddings partially address this issue (some labels are "Liberal" or "Conservative"), but labels are not available for all views. Going forward, we hope to classify the viewpoint of the original claim and then generate a claim with a desired orientation.

Finally, improvements to the model may include retrieval-based approaches to handle low-frequency terms and named entities, as sequence-to-sequence models are likely to have trouble in this environment. One possibility is to incorporate external knowledge with entity linking over Wikipedia articles to find semantically-relevant substitutions.

Chapter 6: Semantic Challenges for Argumentation

In Section 4.1, we examined the role of the *intrinsic* semantic properties of claims and premises. However, an alternate, complementary view of semantics involves the *contextual* properties of arguments, which necessitates an understanding of the relations between propositions. While Section 4.1 used the theory of Toulmin [1958] and Aristotle et al. [1954] to annotate and analyze the role of logos at a coarse-grained level, here we examine semantic relations at a fine-grained level by further distinguishing them according to their characterization by argument schemes [Walton, 1995] or acceptability judgments [Freeman, 2000].

Here, we consider two aspects of semantic relations: causality and veracity. Causal relations align with the causal mechanism argument scheme described in the work of Musi et al. [2018] and leveraged in the work of Reisert et al. [2015] and Yanase et al. [2015], where a causal relation may be within or between argumentative discourse units (ADUs). Fact-checking, on the other hand, may be necessary for verifying non-experiential propositions (using the acceptability criteria of Park et al. [2015] or Hua and Wang [2017]) by verifying their content with an authoritative source. Because causality and veracity lie on different dimensions, an argumentative proposition may then both contain a causal relation and be verifiable. Consider the following examples:

1. In a monopsony[1] environment, a minimum wage could actually **increase** employment and wages.

2. Nevada should have a "no confidence" vote in elections to **increase** turnout.

3. Donald Trump was the first politician to believe in a minimum wage.

The first example is referential, where the author is appealing to the credibility of research in economics. While this sentence contains the causal verb *"increase,"* fact-checking would not be

[1] https://en.wikipedia.org/wiki/Monopsony

applicable. However, the second sentence both contains a causal relation and a verifiable proposition. A reasonable counter to this proposition is that Nevada has had the choice *"None of these options"* since 1975. Finally, the third example is verifiable but does not contain a causal relation.

We thus develop methods both for detecting when a causal relation is present [Hidey and McKeown, 2016] and when a veracity relation is present between a verifiable proposition and evidence [Hidey and Diab, 2018]. These techniques allow us to obtain claims and evidence that can be used in a *retrieval-based* approach to argument generation, where the focus is on the procurement of argumentative content, in contrast with the *generation-based* approach of Section 5.2, where the focus is on modification. Our methods for causal relation detection can be used to identify causal reasoning given a candidate set of claims and premises. Similarly, evidence retrieved using our fact-checking system can be used as premises that support or attack another proposition. We can then use methods such as our approach in Section 4.2 to identify the presence of a relation between propositions and create argumentative structures. These structures can then can be incorporated into a system such as our hybrid retrieval/generation-based approach described in Section 5.1. While we do not create full argumentative structures in our work, we do examine how to use causal relations to retrieve claims in a specialized domain (household electricity consumption) and fact check those claims for veracity.

As discussed in Section 2.3.2, detecting causal relations is difficult without explicit lexical markers. In our work, we develop an approach to discover new explicit markers for causality and improve causal relation detection when those markers are present (Section 6.1). Also, in line with the adversarial approaches discussed in Section 2.3.1 that illustrate the difficulties of natural language inference tasks, we present an adversarial dataset consisting of real-world challenges for fact-checking and an initial approach towards solving the problems we introduced (Section 6.2). Finally, we present an application that considers both types of semantic relations - a system that uses causal relation detection to mine suggestions from social media for reducing electricity usage and fact checks the retrieved suggestions (Section 6.3).

6.1 Causal Reasoning

While the methods discussed in Sections 5.1 and 5.2 provide full arguments and counter-claims, we can also retrieve propositions with causal relations to use as claims or premises (as supporting or refuting evidence), as in the work of Reisert et al. [2015] and Yanase et al. [2015]. Previous work in labeling argument schemes (or inference rules) to link premises to claims found that causal reasoning is the most prevalent in a corpus of essays [Musi et al., 2018]. We thus hypothesize that causal relations are a valuable source of information to detect implicit and explicit inference rules when retrieving arguments.

Causal relations may occur either explicitly with lexical indicators or implicitly [Prasad et al., 2008]. Implicit causal relation detection remains difficult [Roze et al., 2019] whereas explicit causal relations are relatively easy to identify [Pitler and Nenkova, 2009]. Consider the following examples:

1. In the classical model of the minimum wage, it is very clear that a minimum wage will **lead to** higher unemployment among low wage workers.

2. In a monopsony environment, a minimum wage could actually **increase** employment and wages.

3. This proposal would increase the minimum wage. Walmart will begin mass layoffs.

These examples indicate a relation between a cause and an expected effect. In the first case, the lexical indicator *"lead to"* occurs within an ADU and the effect attacks the idea that we should increase the minimum wage. On the other hand, for the second case *"increase"* is part of the effect but also indicates a causal relation and it supports raising the minimum wage. These statements could be used to dispute and support, respectively, a minimum wage increase, and have opposing stances. Finally, for the third case, there is no explicit lexical indicator, but the stated effects of increasing the minimum wage are the actions of Walmart. In this case the implicit causal relation links two ADUs and requires world knowledge and inference (and perhaps an implicit warrant

[Habernal et al., 2018a]).

In the first two cases, the causal relation is presented explicitly, making these cases easier to predict than the third case when the causal relation is implicit. We can thus leverage a model for explicit causal relations to find implicit supporting and attacking arguments as a **retrieval-based** approach to argument generation, where we identify candidate arguments rather than generate them. In our work, we build a new corpus and model for causal relation detection given the limitations of existing resources for explicit causal relations [Hidey and McKeown, 2016].

The PDTB [Prasad et al., 2008] is one such resource for discourse relations, which indicate the semantic connection between spans of text. Causality is one type of discourse relation and can be marked explicitly or conveyed implicitly. In the PDTB, there are 102 known explicit discourse markers such as *"and," "but," "after," "in contrast,"* or *"in addition."* Of these, 28 explicitly mark causal relations (e.g., *"because," "as a result," "consequently."*) Even when causality is explicit, there is a wide variety in how it is expressed. In addition to explicit markers, PDTB researchers recognize the existence of an open class of markers, which they call *AltLex*. There is a tremendous amount of variation in how AltLexes are expressed and so the set of AltLexes is arguably infinite in size. In the PDTB, non-causal AltLexes include *"That compares with"* and *"In any event."* Causal AltLexes include *"This may help explain why"* and *"This activity produced."* While implicit relations are very difficult to identify, they are more common than their easily-identified explicit counterparts. AltLexes, on the other hand, fall in the middle; their linguistic variety makes them difficult to identify but their presence improves the identification of causality.

One issue with causality identification is the lack of data. Unsupervised identification on open domain data yields low precision [Do et al., 2011] and while supervised methods on the PDTB have improved [Roze et al., 2019], creating enough labeled data is difficult. Here, we present a distant-labeling approach to causality identification that uses parallel data to identify new causal connectives given a seed set. Our novel approach uses AltLexes that were automatically identified using semi-supervised learning over a parallel corpus. Since we do not know *a priori* what these phrases are, we used a monolingual parallel corpus to identify new phrases that are aligned with known causal connectives. As large corpora of this type are rare, we used Simple and English Wikipedia to create one.

We discuss the limitations of the PDTB and the linguistic background for our hypothesis in Section 6.1.1. We then describe how we leveraged parallel data in English and Wikipedia articles to **distantly label sentences** with causal connectives (Section 6.1.2). We finally **train a classifier** on this data and self-train to obtain new data (Section 6.1.3), showing that our approach improves over several strong baselines (Section 6.1.4).

6.1.1 Linguistic Background

One disadvantage of the PDTB is that the marked AltLexes are limited only to discourse relations across sentences. We know that there are additional phrases that indicate causality within sentences, but these phrases are neither found in the set of explicit connectives nor AltLexes. Thus, we expand our definition of AltLex to include these markers when they occur within a sentence. Although some phrases or words could be identified by consulting a thesaurus or the Penn Paraphrase Database [Ganitkevitch et al., 2013], we still need the context of the phrase to identify causality.

We hypothesize that there is significant linguistic variety in causal AltLexes. In the set of known explicit connectives there are adjectives (*"subsequent"*), adverbs (*"consequently"*), and prepositions and prepositional phrases (*"as a result"*). We believe that these parts of speech and syntactic classes can be found in AltLexes as well. In addition, verbs and nouns often indicate causality but are not considered explicit connectives.

Some obvious cases of AltLexes are the verbal forms of connectives such as *"cause"* and *"result."* In addition to these verbs, there exist other verbs that can occur in causal contexts but are ambiguous. Consider that *"make"* and *"force"* can replace *"cause"* in this context:

The explosion **made** people evacuate the building.

The explosion **forced** people to evacuate the building.

The explosion **caused** people to evacuate the building.

However, the words can not be substituted in the following sentence:

The baker **made** a cake.

*The baker **caused** a cake.

*The baker **forced** a cake.

Furthermore, verbs such as *"given"* may replace additional causal markers:

It's not surprising he is tired **since** he did not get any sleep.

It's not surprising he is tired **given that** he did not get any sleep.

There are also some phrases with the same structure as partial prepositional phrases like *"as a result"* or *"as a result of,"* where the pattern is a preposition and noun phrase followed by an optional preposition. Some examples of these phrases include *"on the basis of," "with the goal of,"* and *"with the idea of."*

We may also see phrases that are only causal when ending in a preposition such as *"thanks to"* or *"owing to." "Lead"* may only be causal as a part of *"lead to"* and the same for *"develop"* versus *"develop from."* In addition, prepositions can affect the direction of the causality. Comparing *"resulting in"* versus *"resulting from,"* the preposition determines that the latter is of the "reason" class and the former is of the "result" class.

Ultimately, we want to be able to detect these phrases automatically and determine whether they are a large/small and open/closed class of markers.

6.1.2 Data

In order to discover new causal connectives, we can leverage existing information about known causal connectives. It should be the case that if a phrase is a causal AltLex, it will occur in some context as a replacement for at least one known explicit connective. Thus, given a large dataset, we would expect to find some pairs of sentences where the words are very similar except for the connective. This approach requires a parallel corpus to identify new AltLexes. As large English paraphrase corpora are rare, we draw from previous work identifying paraphrase pairs between English and Simple Wikipedia [Hwang et al., 2015]. As Simple Wikipedia has been shown to be not just a simplified version of English Wikipedia [Yasseri et al., 2012], we can mine paraphrase pairs of similar language complexity.

The dataset we used was created from the English and Simple Wikipedias from September 11, 2015. We used the software WikiExtractor[2] to convert the XML into plain text. All articles with the same title were paired and any extra articles were ignored. Each article was lemmatized, parsed (both constituent and dependency), and named-entity tagged using the Stanford CoreNLP suite [Manning et al., 2014]. We wish to identify paraphrase pairs where one element is in English Wikipedia and one is in Simple Wikipedia. Furthermore, we do not limit these elements to be single sentences because an AltLex can occur within a sentence or across sentences.

Previous work [Hwang et al., 2015] created a score for similarity (WikNet) between English Wikipedia and Simple Wikipedia. Many similarity scores are of the following form comparing sentences W and W':

$$s(W, W') = \frac{1}{Z} \sum_{w \in W} \max_{w' \in W'} \sigma(w, w') idf(w) \tag{6.1}$$

where $\sigma(w, w')$ is a score[3] between 2 words and Z is a normalizer ensuring the score is between 0 and 1. For their work, they created a score where $\sigma(w, w') = \sigma_{wk}(w, w') + \sigma_{wk}(h, h')\sigma_r(r, r')$. σ_{wk} is a distance function derived from Wiktionary by creating a graph based on words appearing in a definition. h and h' are the governors of w and w' in a dependency parse and r and r' are the relation. Similar sentences should have similar structure and the governors of two words in different sentences should also be similar. σ_r is 0.5 if h and h' have the same relation and 0 otherwise.

For this work, we also include partial matches, as we only need the connective and the im-

[2] `http://attardi.github.io/wikiextractor/`

[3] The score is not a metric, as it is not symmetric.

Method	Max F1
WikNet	0.4850
WikNet, λ = 0.75	0.5981
Doc2Vec	0.6226
Combined	0.6263

Table 6.1: Paraphrase Results

mediate surrounding context on both sides. If one sentence contains an additional clause, it does not affect whether it contains a connective. Thus, one disadvantage to this score is that when determining whether a sentence is a partial match to a longer sentence or a shorter sentence, the longer sentence will often be higher as there is no penalty for unmatched words between the two elements. We experimented with penalizing content words that do not match any element in the other sentence. The modified score, where W and W' are nouns, verbs, adjectives, or adverbs, is then:

$$\begin{aligned} s(W, W') = \frac{1}{Z} \sum_{w \in W} \max_{w' \in W'} \sigma(w, w') idf(w) \\ -\lambda(|W' - W| + |W - W'|) \end{aligned} \tag{6.2}$$

We also compared results with a model trained using doc2vec [Le and Mikolov, 2014] on each sentence and sentence pair and identifying paraphrases with their cosine similarity.

As these methods are unsupervised, only a small amount of annotated data is needed to tune the similarity thresholds. Two graduate computer science students annotated a total of 45 Simple/English article pairs. There are 3,891 total sentences in the English articles and 794 total sentences in the Simple Wikipedia articles. Inter-annotator agreement (IAA) was 0.9626, computed on five of the article pairs using Cohen's Kappa. We tune the threshold for each possible score: for doc2vec the cosine similarity and for WikNet the scoring function. We also tune the lambda penalty for WikNet. F1 scores were calculated via grid search over these parameters and the best settings are a combined score using doc2vec and penalized WikNet with $\lambda = 0.75$ where a pair is considered to be a paraphrase if either threshold is greater than 0.69 or 0.65 respectively.

Using the combined score we obtain 187,590 paraphrase pairs. After combining and deduping this dataset with the publicly available dataset released by [Hwang et al., 2015], we obtain 265,627 pairs, about 6 times as large as the PDTB. An example paraphrase pair is provided:

> Bleeding Gums Murphy appears to Lisa in a cloud near the end of the episode with Darth Vader, Mufasa, and James Earl Jones saying "This is CNN."
>
> When a deceased Bleeding Gums Murphy appears to Lisa in a cloud near the end he is joined by Darth Vader, Mufasa, and James Earl Jones saying "This is CNN."

However, these examples do not include a causal relation, so we also need a way to automatically label data. To use this dataset for training a model to distinguish between causal and non-causal instances, we use the paired data to identify pairs where an explicit connective appears in at least one element of the pair. The explicit connective can appear in a Simple Wikipedia sentence or an English Wikipedia sentence. We then use patterns to find new phrases that align with these connectives in the matching sentence.

To identify a set of seed words that unambiguously identify causal and non-causal phrases we examine the PDTB. As seen in Table 6.2, causal relations fall under the Contingency class and Cause type. We consider connectives from the PDTB that either only or never appear as that type. The connective *"because"* is the only connective to be almost always a "reason" connective, whereas there are 11 unambiguous connectives for "result" - *"accordingly," "as a consequence," "as a result," "consequently," "hence," "in response," "so that," "subsequently," "thereby," "therefore,"* and *"thus."* "So" when used as a subordinating conjunction is also unambiguously causal. There were many markers that were unambiguously not causal (e.g. *"but," "though," "still," "in addition."*)

In order to label paraphrase data, we use constraints to identify possible AltLexes.[4] We used Moses [Koehn et al., 2007] to train an alignment model on the created paraphrase dataset. Then for every paraphrase pair we identify any connectives that match with any potential AltLexes. Based on our linguistic analysis, we require these phrases to contain at least one content word, which

[4]We do not attempt to label arguments at this point.

Class	Type	Subtype
Temporal		
Contingency	Cause	reason
		result
	Pragmatic cause	
	Condition	
	Pragmatic condition	
Comparison		
Expansion		

Table 6.2: PDTB Discourse Classes

we identify based on part of speech. We also draw on previous work [Pitler and Nenkova, 2009] that used the left and right sibling of a phrase. Therefore, we use the following rules to label new AltLexes:

1. Must be less than 7 words.
2. Must contain at least one content word:
   (a) A non-proper noun
   (b) A non-modal and non-auxiliary verb
   (c) An adjective or adverb
3. Left sibling of the connective must be a noun phrase, verb phrase, or sentence.
4. Right sibling of the connective must be a noun phrase, verb phrase, or sentence.
5. May not contain a modal or auxiliary verb.

Because connectives identify causality between events or agents, we require that each potential connective link two events/agents. We define an event or agent as a noun, verb, or an entire sentence. This means that we require the left sibling of the first word in a phrase and the right sibling of the last word in a phrase to be an event, where a sibling is the node at the same level in the constituent parse. We also require the left and right sibling rule for the explicit connectives, but we allow additional non-content words (for example, we would mark "because of" as a connective rather than "because." We then mark the AltLex as causal or not causal.

Example	Pair
1	Both flora and fauna are scarce **because of** the harsh climate.
	Both flora and fauna are scarce **owing to** the harsh climate.
2	Around the time of the fight, Punk and Dinero had stopped showing up on TNA shows, which lead to speculation that he was fired **because of** the incident.
	Around the time of the scuffle, Punk and Dinero stopped appearing on TNA shows, leading to speculation he was fired **for** the incident.
3	1946 was the first year in which the trustees selected works for entry, **rather than** displaying all those entered.
	In 1946 the trustees selected works for entry, **instead of** displaying all the entries.

Table 6.3: Examples of Aligned AltLexes from our Distant-Labeled Dataset

Given that the paraphrases and word alignments are noisy, we use the syntactic rules to decrease the amount of noise in the data by more precisely determining phrase boundaries. These rules are the same features used by Pitler and Nenkova [2009] for the early work on the PDTB on explicit connectives. These features were successful on the Wall Street Journal and they are applicable for other corpora as well. Also, they are highly indicative of discourse/non-discourse usage so we believe that we are improving on noisy alignments without losing valuable data. In the future, however, we would certainly like to move away from encoding these constraints using a rule-based method and use a machine learning approach to automatically induce rules.

This method yields 72,135 non-causal and 9,190 causal training examples. We present causal and non-causal examples resulting from this process in Table 6.3, with the aligned AltLex highlighted in boldface. In Example 1 (a causal case), the explicit unambiguous connective *"because of"* aligns with the newly-discovered connective *"owing to,"* which is not in the PDTB but is almost unambiguously causal.[5] In another causal instance (Example 2), *"because of"* aligns with the ambiguous connective *"for"* and providing us with a datapoint for disambiguating difficult connectives. Finally, Example 3 shows a non-causal case where two different contrastive AltLexes are aligned.

Although the dataset is noisy, it is larger than the PDTB and was derived automatically. There are 35,136 argument pairs in the PDTB marked with one of the three relations that implies a

[5] Although not entirely unambiguous, as one could utter a sentence like *"That's the money I'm still owing to John,"* the phrase is defined as meaning *"because of"* in several dictionaries.

discourse connective (Implicit, Explicit, and AltLex), and of these 6,289 are causal. Of the 6,289 causal pairs, 2,099 are explicit and 273 contain an AltLex.

6.1.3 Methods

Given training data labeled by this distant supervision technique, we can now treat this problem as a supervised learning problem and create a classifier to identify causality.

We consider two classes of features: features derived from the **parallel corpus** data and **lexical semantic** features. The parallel corpus features are created based on where AltLexes are used as paraphrases for causal indicators and in what context. The lexical semantic features use FrameNet, WordNet, and VerbNet to derive features from all the text in the sentence pair.[6] These lexical resources exploit different perspectives on the data in complementary ways. The parallel corpus features encourage the classifier to select examples with AltLexes that are likely to be causal whereas the lexical semantic features allow the classifier to consider context for disambiguation.

Parallel Corpus Features We create a set of features from the parallel corpus based on a KL-divergence score to encourage the identification of phrases that replace causal connectives. Consider the following datapoints and assume that they are aligned in the parallel corpus:

> I was late **because of** traffic.
>
> I was late **due to** traffic.

We want both of these examples to have a high score for causality because they are interchangeable causal phrases. Similarly, we want non-causal phrases that are often aligned to have a high score for non-causality.

We define several distributions in order to determine whether an AltLex is likely to replace a known causal or non-causal connective. We consider all aligned phrases, not just ones containing a

[6] We explored FrameNet due to the usage in prior work [Riaz and Girju, 2013], while other resources such as PropBank [Kingsbury and Palmer, 2002] and AMR [Banarescu et al., 2013] could be investigated in future work.

causal or non-causal connective to attempt to reduce noisy matches. The idea is that non-connective paraphrases will occur often and in other contexts.

The following conditional Bernoulli distributions are calculated for every aligned phrase in the dataset, where w is the phrase, s is the sentence it occurs in, c is "causal" and nc is "not causal":

$$p_1 = p(w_1 \in s_1 | rel(s_1) \in \{c\}, w_1 \notin s_2) \tag{6.3}$$

$$p_2 = p(w_1 \in s_1 | rel(s_1) \in \{nc\}, w_1 \notin s_2) \tag{6.4}$$

We compare these two distributions to other distributions with the same word and in a different context (where o represents "other"):

$$q_1 = p(w_1 \in s_1 | rel(s_1) \in \{nc, o\}, w_1 \notin s_2) \tag{6.5}$$

$$q_2 = p(w_1 \in s_1 | rel(s_1) \in \{c, o\}, w_1 \notin s_2) \tag{6.6}$$

We then calculate $D_{KL}(p_1||q_1)$ and $D_{KL}(p_2||q_2)$. In order to use KL-divergence as a feature, we multiply the score by $(-1)^{p<q}$ and add a feature for **causal** and one for **non-causal**.

Lexical Semantic Features As representations of events are composed of predicates and arguments often referred to by nouns and verbs, we consider using lexical semantic resources that have defined hierarchies for nouns and verbs. We thus use the lexical resources FrameNet, WordNet, and VerbNet as complementary resources from which to derive features for words both inside and outside the AltLex, disambiguating their context using part-of-speech tags or dependency parse trees. We hypothesize that these semantic features provide context not present in the text; from these we are able to infer causal and anti-causal properties.

FrameNet is a resource for frame semantics, defining how events, objects, and relations interact, and provides an annotated corpus of English sentences. WordNet provides a hierarchy of word senses and we show that the top-level class of verbs is useful for indicating causality. VerbNet provides a more fine-grained approach to verb categorization that complements the views provided by FrameNet and WordNet.

In **FrameNet**, a semantic frame is a conceptual construction describing events or relations and their participants [Ruppenhofer et al., 2006]. Frame semantics abstracts away from specific utterances and ordering of words in order to represent events at a higher level. While in Section 4.3.2, we leveraged predicted frames to represent social events such as agreement, here we use features derived from the FrameNet resource to represent causal events. There are over 1,200 semantic frames in FrameNet and some of these can be used as evidence or counter-evidence for causality. In the work of Riaz and Girju [2013], they identified 18 frames as causal (e.g. "Purpose," "Internal cause," "Reason," "Trigger").

We use these same frames to create a lexical score based on the FrameNet 1.5 corpus. This corpus contains 170,000 sentences manually annotated with frames. We used a part-of-speech-tagged version of the FrameNet corpus and for each word and tag, we count how often it occurs with one of the given frames. We only considered nouns, verbs, adjectives, and adverbs. We then calculate $p_w(c|t)$ and c_{wct}, the probability that a word w is causal given its part-of-speech tag t and its count, respectively. The lexical score of a word i is calculated by using the assigned part-of-speech tag and is given by $CS_i = p_{w_i}(c|t_i) \log c_{w_i c t_i}$. The total score of a sequence of words is then $\sum_{i=0}^{n} CS_i$.

We also took this further and determined what frames are likely to be *anti-causal*. We started with a small set of seed words derived directly from 11 discourse classes (types and subtypes from Table 6.2), such as "Compare," "Contrast," "Explain," "Concede," and "List." We expanded this list using WordNet synonyms for the seed words. We then extracted every frame associated with their stems in the stemmed FrameNet corpus. These derived frames were manually examined to develop a list of 48 anti-causal frames, including "Statement," "Occasion," "Relative time," "Evidence," and "Explaining the facts."

We create an anti-causal score using the FrameNet corpus just as we did for the causal score.

The total anti-causal score of a sequence of words is $\sum_{i=0}^{n} ACS_i$ where $ACS_i = p_{w_i}(a|t_i) \log c_{w_i a t_i}$ for anti-causal probabilities and counts. We split each example into three parts: the text before the AltLex, the AltLex, and the text after. Each section is given a causal score and an anti-causal score. Overall, there are six features derived using FrameNet: causal score and anti-causal score for each part of the example.

In **WordNet**, words are grouped into "synsets," which represent all synonyms of a particular word sense. Each word sense in the WordNet hierarchy has a top-level category based on part of speech [Miller, 1992]. Every word sense tagged as noun, verb, adjective, or adverb is categorized. Some examples of categories are "change," "stative," or "communication." We only include the top level because of the polysemous nature of WordNet synsets. We theorize that words having to do with change or state should be causal indicators and words for communication or emotion may be anti-causal indicators.

Similar to the FrameNet features, we split the example into three sections. However, we also consider the dependency parse of the data. We believe that causal relations are between events and agents that are represented by nouns and verbs. Events can also be represented by predicates and their arguments, which is captured by the dependency parse. As the root of a dependency parse is often a verb and sometimes a noun or adjective, we consider the category of the root of a dependency parse and its arguments.

We include a categorical feature indicating the top-level category of the root of each of the three sections, including the AltLex. For both sides of the AltLex, we include the top-level category of all arguments as well. If a noun has no category, we mark it using its named-entity tag. If there is still no tag, we mark the category as "none."

VerbNet is a resource devoted to storing information for verbs [Schuler et al., 2000]. In contrast to WordNet, VerbNet provides a more fine-grained description of events while focusing less on polysemy. Some examples of VerbNet classes are "force," "indicate," and "wish." In VerbNet, there are 273 verb classes, and we include their presence as a categorical feature. Similar to WordNet, we use VerbNet categories for three sections of the sentence: the text pre-AltLex, the

AltLex, and the text post-AltLex. Unlike WordNet, we only mark the verbs in the AltLex, root, or arguments.

Finally, we consider **interactions** between the WordNet and VerbNet features. As previous work [Marcu and Echihabi, 2002, Biran and McKeown, 2013] used word pairs successfully, we hypothesize that pairs of higher-level categories will improve classification without being penalized as heavily by the sparsity of dealing with individual words. Thus we include interaction features between every categorical feature for the pre-AltLex text and every feature for the post-AltLex text.

In all, we include the following features (*L* refers to the AltLex, *B* refers to the text *before* the AltLex and *A* refers to the text *after* the AltLex):

1. FrameNet causal score for L, B, and A.
2. FrameNet anti-causal score for L, B, and A.
3. WordNet top-level of L.
4. WordNet top-level of the root of B and A.
5. WordNet top-level for arguments of B and A.
6. VerbNet category for verb at the root of L.
7. VerbNet top-level category for any verb in the root of B and A.
8. VerbNet top-level category for any verbs in the arguments of B and A.
9. Categorical interaction features between the features from B and the features from A.

6.1.4 Results

We evaluated our methods on two manually annotated test sets. We used one of these test sets for development only. For this set, one graduate computer science student and two students from the English department annotated a set of Wikipedia articles by marking any phrases they considered to indicate a causal relationship and marking the phrase as REASON or RESULT. Wikipedia articles from the following categories were chosen as we believe they are more likely to contain causal relationships: science, medicine, disasters, history, television, and film. For each article in this category, both the English and Simple Wikipedia articles were annotated. A total of 12 article pairs were annotated. IAA was computed to be 0.31 on two article pairs using Kripendorff's alpha.

IAA was very low and we also noticed that annotators seemed to miss sentences containing causal connectives. It is easy for an annotator to overlook a causal relation when reading through a large quantity of text. Thus, we created a new task that required labeling a connective as REASON, RESULT, or NON-CAUSAL when provided with the sentence containing the connective. For testing, we used CrowdFlower to annotate the output of the system using this method. We created a test set by annotating 600 examples, where the system labeled 300 as causal and 300 as non-causal, selected from the same set of Wikipedia articles and ensuring there was no overlap with our distant-labeled training data. Contributors were limited to the highest level of quality and from English-speaking countries. We required 7 annotators for each data point. The IAA was computed on the qualification task that all annotators were required to complete. There were 15 questions on this task and 410 annotators. On this simplified task, the IAA improved to 0.69.

We also considered evaluating the results on the PDTB but encountered several issues. As the PDTB only has a limited set of explicit intra-sentence connectives marked, this would not show the full strength of our method. Many causal connectives that we discovered are not annotated in the PDTB. Alternatively, we considered evaluating on the AltLexes in the PDTB but these examples are only limited to inter-sentence cases, whereas the vast majority of our automatically annotated training data was for the intra-sentence case. Thus we concluded that any evaluation on the PDTB would require additional annotation. Our goal in this work was to identify new ways in which causality is expressed, unlike the PDTB where annotators were given a list of connectives and asked to determine discourse relations.

	Accuracy	True Precision	True Recall	True F-measure
Most Common Class	63.50	60.32	82.96	69.85
$CONN$	62.21	78.47	35.64	49.02
LS	67.68	61.98	58.51	60.19
KLD	58.03	**91.17**	19.55	32.20
$LS \cup KLD$	**73.95**	80.63	64.35	71.57
$LS \cup LS_{inter}$	72.99	78.54	64.66	70.93
$KLD \cup LS \cup LS_{inter}$	70.09	76.95	58.99	66.78
$LS \cup KLD \cup CONN$	71.86	70.28	**77.60**	**73.76**
$Bootstrapping_1$	79.26	**77.97**	82.64	80.24
$Bootstrapping_2$	**79.58**	77.29	**84.85**	**80.90**

Table 6.4: Experimental Results

Most Common Class - assign the label to the class it occurs most often in the training data
$CONN$ - connective as a feature
LS - full set of lexical semantic features described in Section 6.1.3
LS_{inter} - interaction between LS described in Section 6.1.3
KLD - KL-divergence features described in Section 6.1.3
$Bootstrapping_n$ - results after n rounds of bootstrapping using the feature set $KLD \cup LS \cup LS_{inter}$

We tested our hypothesis by training a binary[7] classifier on our data using the full set of features we just described. We used a linear Support Vector Machine (SVM) classifier [Vapnik, 1998] trained using stochastic gradient descent (SGD) through the sci-kit learn package. [Garreta and Moncecchi, 2013][8] We used elasticnet to encourage sparsity and tuned the regularization constant α through grid search.

We use two baselines. The first baseline is the most common class of each AltLex according to its class in the initial training set. For example, *"caused by"* is almost always a causal AltLex. A second baseline uses the AltLex itself as a categorical feature and is shown as $CONN$ in Table 6.4. For comparison, this is the same baseline used by Pitler and Nenkova [2009] on the explicit discourse relations in the PDTB. We compare these two baselines to ablated versions of our system. We evaluate on the KL-divergence (KLD) and semantic (LS and LS_{inter}) features described in Section 6.1.3. LS consists of features 1-8, all the FrameNet, VerbNet, and WordNet features. LS_{inter} includes only the interaction between categorical features from WordNet and VerbNet.

[7]We combine REASON and RESULT into one CAUSAL class but future work could distinguish between REASON, RESULT, or NON-CAUSAL.

[8]We also considered a logistic regression classifier.

We calculate accuracy and true precision, recall, and F-measure for the causal class. As seen in Table 6.4, the best system ($LS \cup KLD \cup CONN$) outperforms the baselines.[9] The lexical semantic features by themselves (LS) are similar to those used by [Riaz and Girju, 2014] although on a different task and with the WordNet and VerbNet features included. Note that the addition of the Altlex words and KL-divergence ($LS \cup KLD \cup CONN$) yields an absolute increase in f-measure of 13.57 points over lexical semantic features alone.

Bootstrapping Our method for labeling AltLexes lends itself naturally to a bootstrapping approach. As we are using explicit connectives to identify new AltLexes, we can also use these new AltLexes to identify additional ones. We consider any unlabeled pairs remaining from the 187,590 paraphrase pairs described in Section 6.1.2 where at least one aligned phrase in the pair contains one of our newly discovered AltLexes. We then use our classifier to automatically label these new data points and remove any phrases where the classifier did not agree on both elements in the pair. The set of features used were the $KLD \cup LS \cup LS_{inter}$ features as these performed best on the development set and it provides us with a connective-independent way of labeling causal relations. We use early stopping on the development data to identify the point when adding additional data is not worthwhile. The bootstrapping method converges quickly. After 2 iterations we see a decrease in the F-measure of the development data.

We then evaluate on the heldout test set created by CrowdFlower workers (which has no overlapping connectives with any bootstrapped data added after two rounds). The increase in performance on the test data is significant. In Table 6.4, $Bootstrapping_n$ refers to results after n rounds of bootstrapping. Bootstrapping yields improvement over the supervised method with an absolute gain of 7.14 points. These results show that the model is able to learn features of causality that are not entirely dependent on the connective itself.

Discussion Of note is that the systems without connectives (combinations of LS, LS_{inter}, and KLD) perform well on the development set without using any lexical features. Using this system

[9] These results are statistically significant by a binomial test with $p < 7 * 10^{-6}$.

	True Precision	True Recall	True F-measure
$FrameNet$	67.88	53.14	59.61
$WordNet$	76.92	9.52	16.94
$VerbNet$	38.70	3.80	6.92

Table 6.5: Semantic Feature Ablation

enables the discovery of new AltLexes during bootstrapping, as we cannot rely on having a closed class of connectives but need a way of classifying connectives not seen in the initial training set.

Also important is that the Altlex by itself ($CONN$) performs poorly. In comparison, in the task of identifying discourse relations in the PDTB these features yield an 75.33 F-score and 85.85% accuracy in distinguishing between discourse and non-discourse usage [Pitler and Nenkova, 2009] and an accuracy of 93.67% when distinguishing between discourse classes. Although this is a different data set, this shows that identifying causality when there is an open class of connectives is much more difficult. We believe the connective by itself performs poorly because of the wide linguistic variation in these alternative lexicalizations. Many connectives appear only once or not at all in the training set, so the additional features are required to improve performance.

In addition, the "most common class" baseline is a strong baseline. The strength of this performance provides some indication of the quality of the training data, as the majority of the time the connective is indicative of its class in the held-out test data. However, the overall accuracy is still much lower than if we use informative features.

The KLD and LS feature sets appear to be complementary. The KLD feature sets have higher precision on a smaller section of the data, whereas the LS system has higher recall overall. These lexical semantic features likely have higher recall because these resources are designed to represent *classes* of words rather than individual words. Some connectives occur very rarely, so it is necessary to generalize the key aspects of the connectives and class-based resources provide this capability.

Feature Ablation In order to determine the contribution of each lexical resource, we perform additional feature ablation for each of FrameNet, WordNet, and VerbNet. As seen in Table 6.5,

Example	Sentence
1	Language is reduced to simple phrases or even single words, eventually **leading to** complete loss of speech.
2	Kulap quickly accelerated north, **prompting** the PAGASA to issue their final advisory on the system.
3	When he finally changed back, Buffy stabbed him **in order to** once again save the world.
4	Agricultural potential is generally poor, due to the natural infertility of soils and **the prevalence of** swamps and lakes left by departing ice sheets, and short growing seasons prohibit all but the hardiest of crops.

Table 6.6: Causal Examples Identified by our Distant-Labeling Approach

the lexical semantic resources each contribute uniquely to the classifier. The FrameNet features provide most of the performance of the classifier. The WordNet and VerbNet features, though not strong individually, supply complementary information and improve the overall performance of the LS system (see Table 6.4) compared to just using FrameNet alone.

Analysis To further understand the benefits of the distant-labeling process, we present examples in Table 6.6 containing causal connectives that were predicted correctly or discovered during bootstrapping. The model ($LS \cup KLD \cup CONN$) correctly identifies some causal relations that neither baseline identifies (Examples 1 and 2). These examples contain non-standard causal connectives (*"leading to"*[10] and *"prompting"*) and occur infrequently in the data, so the lexical semantic features help to identify them. After two rounds of bootstrapping, the system is able to recover additional examples that were not found previously such as Examples 3 and 4. These connectives occur rarely or not at all in the initial training data and are only recovered because of the improvements in the model.

In comparison, the bootstrapping process also labels several ambiguous known connectives that may allow the substitution of an unambiguous connective. Table 6.7 provides pairs of examples where one sentence contains an ambiguous connective and other contains an unambiguous connective. For instance, Example 1 shows one sentence with a causal relation expressed using

[10]The phrase is not entirely unambiguous, as one could refer to *"the door leading to the next room,"* but multiple dictionaries define *"lead to"* as *"causing something to happen or exist."*

Example	Pair
1	As a result of their fast growth rate, antlers are considered a handicap since there is an incredible nutritional demand on deer to re-grow antlers annually, **and** can be honest signals of metabolic efficiency and food gathering capability.
	To re-grow antlers each year uses up nutrition, **so** they are honest signals of food gathering capability.
2	A Common Blackbird has an average life expectancy of 2.4 years, **and**, based on data from bird ringing, the oldest recorded age is 21 years and 10 months.
	On average, Blackbirds live to be 2.4 years old, **but** some have been found to be 20 years old.

Table 6.7: Ambiguous/Unambiguous Pairs Identified by the Bootstrapping Approach

the connective *"and"* and another with the unambiguous causal connective *"so"* used for distant-labeling. On the other hand, the connective *"and"* is often non-causal, as seen in Example 2, which was identified by bootstrapping using its corresponding non-causal connective *"but"*.

6.1.5 Conclusions, Limitations, and Future Work

We have provided a method for automatically building a training set for causality and identifying and classifying phrases that indicate the presence of a causal relation. We demonstrated statistically significant improvement using our semantic and parallel corpus features over strong-performing baselines for explicit discourse relation detection; the text in the AltLex alone is not sufficient to accurately identify causality. We also showed via ablation studies and qualitative analysis that our features are informative by themselves and perform well even on rarely occurring examples.

Ultimately, the focus of this work is to improve detection of causal relations. Thus, we did not conduct some experiments, such as an evaluation of the quality of the automatically annotated corpus or the model performance on the three-way task (REASON, RESULT, or NON-CAUSAL). Our use of distant supervision demonstrates that we can use a large amount of possibly noisy data to develop an accurate classifier. To evaluate the intermediate step would have required an additional annotation process. Future work can examine how to automatically identify the span of AltLexes and causal relations in a manner similar to how the PDTB is annotated [Prasad et al., 2008]. Additionally, given that we have distant-labeled data and annotations for the fine-grained distinction of causality between REASON and RESULT, future work should include better modeling approaches such as neural methods that can determine the direction of the relation. Understanding the difference between REASON and RESULT is key for using causal relations in a downstream argumentative task.

Although we have focused exclusively on Wikipedia, these methods could be adapted to other domains such as Reddit. Causality is not easily expressed in English using a fixed, contiguous set of phrases [Dunietz et al., 2017], so we would expect these methods to apply to formal and informal text ranging from news and journals to social media.

Finally, future work could incorporate causal relation detection into our hybrid argument generation system described in Section 5.1.2. Retrieval of supporting or refuting evidence in the form of causal relations could be done with a combination of Google search over Wikipedia articles and our classifier. However, future work would also need to predict the stance of a proposition containing a causal relation in order to incorporate the evidence into a counter-argument. While the IBM Debater project has data and models for context-dependent evidence detection using Wikipedia [Bar-Haim et al., 2017a,b], future work would need to determine how well these systems perform specifically on causal relations. Furthermore, as discussed, the ability to distinguish between REASON and RESULT is necessary to determine what a candidate argument is supporting or refuting.

6.2 Fact-Checking

While the causal relations discussed in Section 6.1 provide a method for identifying the type of reasoning used in an argument and for retrieving arguments using a causal argument scheme, a complementary approach involves the verification of propositions. Given a verifiable, "check-worthy" proposition [Freeman, 2000, Park and Cardie, 2014, 2018, Hassan et al., 2017, Hua and Wang, 2017], an ideal property of an argument generation/retrieval system is the ability to fact check the proposition using an authoritative source. As noted in Section 4.1, relations between propositions may be supporting/agreeing or attacking/disagreeing. The ability to provide evidence and cite sources for retrieved verifiable arguments allows the model to handle support relations as well as attack relations against one's own arguments using a prebuttal approach. Likewise, the undercutting or rebuttal of an opponent's arguments may require the retrieval of contradictory evidence. The following claim is taken from Change My View along with a proposition from the response:

- Countries should have a "no confidence" vote in elections if they want to increase turnout.
- The US state of Nevada has had a choice called "none of these candidates" since 1975.

In this example, the claim of the original poster is rebutted by directly challenging the truth of the original claim and providing supporting evidence from Wikipedia. The argument of an opponent may contain incorrect statements and finding factual evidence to refute these claims is likely to help when generating a counter-argument. As in this case, per our goal of identifying and retrieving effective counter-arguments, fact-checking models can provide us with new content that can be used as input to multi-argument fusion.

In order to build effective models for fact-checking, we need models with an understanding of semantic relations. Predicting a veracity relation between a proposition and an external source requires an understanding of contrast. Furthermore, if the external source is not provided, the evidence needs to be obtained from an authoritative source, which may be difficult to find in an era of polarized opinions. Verifying claims using textual sources is therefore a difficult problem, as it requires natural language inference as well as information retrieval. Fact-checking arguments is then a task that jointly benefits from not only asserting or disputing the veracity of a claim but also finding evidence for that position.

In our work, we evaluate the current capabilities for both evidence selection and veracity prediction and build methods that improve on the current state of the art. While semantic relations such as causality are known to be difficult, recent work in adversarial datasets has shown that natural language inference tasks often fail on seemingly simple variations of known problems. Similarly, we develop an adversarial dataset specifically for fact-checking that addresses known

Claim: The Rodney King riots took place in the most populous county in the USA.
Evidence: [Los Angeles Riots] The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.
Evidence: [Los Angeles County] Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.
Label: SUPPORTS

Figure 6.1: Example from FEVER 1.0 Dataset.

challenges in real-world claims found in sources such as news editorials and online debates. Then, we build a system that attempts to address the problems we identified.

We assume access to a trusted resource such as Wikipedia. Wikipedia is one of the few highly regarded sources; according to a recent poll, 64% of British people found Wikipedia to be trustworthy, a larger percentage than even the BBC (61%) [Jordan, 2014]. We also assume that all information can be verified in a subset of sentences in a set of Wikipedia articles. Specifically, we use the FEVER 1.0 shared task dataset [Thorne et al., 2018], which has claims that have been labeled with a veracity label and corresponding evidence from Wikipedia. While prior work has examined community forums [Thorne and Vlachos, 2019] and fact-checking websites such as PolitiFact [Wang, 2017, Alhindi et al., 2018], the FEVER dataset allows for the development of end-to-end fact-checking systems requiring document retrieval and evidence extraction to verify a predicted veracity label (SUPPORTS, REFUTES, NOT ENOUGH INFO). This scenario allows us to conduct experiments in a controlled environment, both for identifying weaknesses in current fact-checking systems and developing approaches towards a solution. A challenging example of the task that requires multi-hop retrieval is given in Figure 6.1. To verify this claim using Wikipedia, one would need to consult two articles: “Los Angeles Riots” to locate the Rodney King riots and “Los Angeles County” to find the most populous US county.

Since the claims in the FEVER 1.0 dataset were manually written using information from Wikipedia, the dataset may lack linguistic challenges that occur in verifying naturally occurring check-worthy claims, such as temporal reasoning or lexical generalizations/specifications. Thorne and Vlachos [2019] designed a second shared task (FEVER 2.0) for participants to create adversarial claims ("attacks") to break state-of-the-art systems and then develop systems to resolve those attacks.

We present a **novel dataset of adversarial examples** for fact extraction and verification in three challenging categories: 1) multiple propositions (claims that require multi-hop document or sentence retrieval); 2) temporal reasoning (date comparisons, ordering of events); and 3) named entity ambiguity and lexical variation (Section 6.2.2). We show that state-of-the-art systems are vulnerable to adversarial attacks from this dataset (Section 6.2.4). In addition, we take steps toward addressing these vulnerabilities, presenting a system for end-to-end fact-checking that brings **two novel contributions using pointer networks**: 1) a document ranking model; and 2) a joint model for evidence sentence selection and veracity relation prediction framed as a sequence labeling task (Section 6.2.3). Our new system achieves state-of-the-art results for FEVER and we present an evaluation of our models including ablation studies (Section 6.2.4).

6.2.1 Problem Formulation and Datasets

The end-to-end fact-checking problem we address is framed in the context of FEVER [Thorne et al., 2018], a task where a system is required to verify a claim by providing evidence from a large collection of Wikipedia articles and predicting whether it is supported (SUPPORTS), refuted (REFUTES), or there is not enough information (NEI) to verify the claim. To be successful, a system needs to predict both the correct label and the correct set of evidence sentences from Wikipedia (for the SUPPORTS and REFUTES cases).

FEVER 1.0 The FEVER 1.0 dataset [Thorne et al., 2018] was created by extracting sentences which contain one or more entities from the top 50,000 most popular Wikipedia pages and altering a sentence in the article to be a new claim — mutating the sentences to be a paraphrase or another edit operation. Then, each claim has been labeled as SUPPORTS, REFUTES, or NEI and paired with the appropriate evidence or the empty set for NEI. Overall, there are 185,445 claims, out of which 90,367 are supported, 40,107 are refuted, and 45,971 are NEI. A training and development set is available from the organizers and the evaluation is performed on a blind test set by submitting predictions to a server. The split in training, development, and test is given in Table 6.8.

FEVER 2.0 Thorne and Vlachos [2019] introduced an adversarial set up for the second FEVER shared task - participants can submit both adversarial claims to break existing systems and a system designed to withstand such attacks. Thorne and Vlachos [2019] provided a baseline set of 1000 adversarial examples using simple but effective entailment-preserving transformations and simple and complex negation. The final dataset for FEVER 2.0 consists of adversarial examples submitted by the participants and is split into a development and a blind test set (Table 6.8). The FEVER 2.0 shared task thus provides the ideal setup to develop novel datasets of claims and evidence that capture some of the challenging problems encountered in verifying naturally-occurring check-worthy claims.

Dataset	Train	Dev.	Blind Test
FEVER 1.0	145,449	19,998	19,998
FEVER 2.0	–	1,174	1,180

Table 6.8: Dataset Statistics

6.2.2 Adversarial Dataset for Fact-checking

While the FEVER dataset is a valuable resource, our goal is to evaluate complex adversarial claims which resemble check-worthy claims found in news articles, speeches, debates, and online discussions. We thus propose three types of attacks to make progress towards this goal: those using multiple propositions, requiring temporal and numerical reasoning, and involving lexical variation.

For the **multi-propositional** type, Graves [2018] notes that professional fact-checking organizations need to synthesize evidence from multiple sources and automated fact-checking systems struggle with claims such as *"Lesotho is the smallest country in Africa."* To verify this claim, an automated system would need to retrieve information about Lesotho such as its geographical area and make multiple comparisons to other African countries. In the FEVER 1.0 dataset, 83.18% of SUPPORTS and REFUTES claims require only a single piece of evidence and 89% require only a single Wikipedia article. Furthermore, our initial analysis on FEVER 1.0 showed that a model can fully retrieve 86% of evidence sentences from Wikipedia where only a single sentence is required, but the number drops to 17% when 2 sentences are required and 3% when 3 or more sentences are required [Hidey and Diab, 2018].

Numerical claims occur frequently in check-worthy claims [Francis and Fact, 2016]. Especially challenging subsets are those containing **temporal reasoning** as identifying temporal relations is a difficult task [Mirza and Tonelli, 2016]. These claims are notable as Rashkin et al. [2017] and Jiang and Wilson [2018] showed that numbers and comparatives are indicative of truthful statements in news; however, the presence of a date alone does not indicate its veracity. In the FEVER 1.0 dataset, 17.81% of the test claims contain dates and only 0.22% contain time information.[11] To understand how current systems perform on these types of claims, we evaluated three state-of-the-art systems from FEVER 1.0 [Hanselowski et al., 2018, Yoneda et al., 2018, Nie et al., 2018], and examined the predictions where the systems disagreed. We found that in characterizing these predictions according to the named entities present in the claims, the most frequent types were numerical and temporal (such as percent, money, quantity, and date).

Finally, adversarial attacks for **lexical variation**, where words may be inserted or replaced or changed with some other edit operation, have been shown to be effective for similar tasks such as natural language inference [Nie et al., 2019] and question answering [Jia and Liang, 2017], so we include these types of attacks as well. For fact-checking, models must be able to match words and entities in a claim with similar words and entities in the evidence to make a veracity prediction. As claims often contain ambiguous entities [Thorne and Vlachos, 2018] or lexical features indicative of credibility [Nakashole and Mitchell, 2014], we desire models resilient to minor changes in entities [Hanselowski et al., 2018] and words [Alzantot et al., 2018].

We thus create an adversarial dataset with examples for each type of attack, with 417 multi-propositional, 313 temporal and 270 lexically variational.

[11] As determined by named entity recognition using Spacy: `https://spacy.io`

Multiple Propositions Check-worthy claims often consist of multiple propositions joined by connectives such as conjunctions. In the FEVER task, this may require retrieving evidence sequentially after resolving entities and events, understanding the role of discourse connectives, and evaluating each proposition.

Consider the claims *"Janet Leigh was from New York."* and *"Janet Leigh was an author."* The Wikipedia page **[Janet Leigh]** contains evidence that she was an author, but makes no mention of New York. So while the latter claim is verifiable, it becomes unverifiable, or NEI, when we conjoin it with the former to make the new claim *"Janet Leigh was from New York and was an author."* We generate claims of the CONJUNCTION type *automatically* by mining claims from the FEVER 1.0 shared task development set and extracting entities from the subject position. We then combine the original claims by replacing the subject in one sentence with a discourse connective such as "and." The new label is SUPPORTS if both original claims are SUPPORTS, REFUTES if one claim is REFUTES and the other is REFUTES or SUPPORTS, and NEI otherwise. While this method results in a bias towards NEI, due to being the most common conjunction type, this provides a fully automated way to generate examples to evaluate model performance.

While CONJUNCTION claims provide a way to evaluate multiple propositions about a single entity, these claims only require evidence from a single page; hence we create new examples requiring reasoning over multiple pages. To create MULTI-HOP examples, we collect claims from FEVER 1.0 and filter those whose evidence obtained from a single Wikipedia page P contains at least one other named entity having a valid Wikipedia page Q. We then modify the existing claim by appending new information about the named entity which can be verified from Q. For example, the claim *"The Nice Guys is a 2016 action comedy film."* from the FEVER 1.0 shared task development set, can be verified using the Wikipedia page **[The Nice Guys]**. The evidence for this claim is a sentence from the article: *"The Nice Guys is a 2016 American neo-noir crime black comedy film directed by Shane Black and written by Black and Anthony Bagarozzi."* We can then make a multi-hop claim by obtaining the Wikipedia page **[Shane Black]** and appending a relative clause from the article to make a new claim: *"The Nice Guys is a 2016 action comedy film* ***directed***

by a Danish screenwriter known for the 1987 action film Lethal Weapon".

While multi-hop retrieval provides a way to evaluate the SUPPORTS and REFUTES cases, composition of multiple propositions may also be necessary to predict the NEI label, as the use of more general or specific phrases may change the relation of the claim to the evidence. This motivated us to add ADDITIONAL UNVERIFIABLE PROPOSITIONS to our claims that change the gold label from S to NEI. We selected claims from FEVER 1.0 development with the SUPPORTS label and added propositions which have no verifiable evidence in Wikipedia. For example, we can add the reduced relative clause *"born in Seattle"*, to make the claim *"Duff McKagan is an American citizen."* unverifiable.

Temporal Reasoning In the context of FEVER, verifying a claim that requires temporal information requires reasoning and comparing dates and times across claims and evidence.

In order to evaluate the ability of current systems to handle temporal reasoning we modify claims from the FEVER 1.0 development set. More specifically, using claims that have the phrase "in <date>" we *automatically* generate seven modified claims using seven simple DATE MANIPULATION heuristics: date arithmetic (addition, subtraction, or in between), date range (before, after, or in between), and date verbalization. For instance, take the claim *"Wildfang was founded in Portland, Oregon in 2001."* This claim can be modified by date subtraction (*"in 2001"* → *"4 years before 2005"*), a date range (*"in 2001"* → *"in the 2000s"*), or date verbalization (*"in 2001"* → *"in the first decade of the twenty-first century"*).

We also create examples requiring MULTI-HOP TEMPORAL REASONING, where the system must evaluate an event in relation to another event. Consider the claim in Figure 6.2. Verification of this claim requires three evidence sentences from two Wikipedia pages. From the page **[William Henry Harrison]**, we can obtain the information that he was the first governor and the date of his death. However, we also need the Wikipedia page for **[Indiana Territory]**. Overall, a system must resolve the entity references (Indiana Territory and its first governor, William Henry Harrison) and compare the dates of the events (the admittance of Indiana in 1816 and the death of Harrison in

Claim: The first governor of the Indiana Territory lived long enough to see it become a state.
Evidence: [William Henry Harrison] Before election as president, Harrison served as the first congressional delegate from the Northwest Territory and the first Governor of Indiana Territory.
Evidence: [William Henry Harrison] However, Harrison died of pneumonia in April 1841, a month after taking office.
Evidence: [Indiana Territory] The Territory of Indiana was an organized incorporated territory of the United States that existed from July 4, 1800, until December 11, 1816, when the remaining southern portion of the territory was admitted to the Union as the state of Indiana.
Label: SUPPORTS

Figure 6.2: Multi-Hop Temporal Reasoning Example

1841). While multi-hop retrieval may resolve references, the model must understand the meaning of *"lived long enough to see"* and evaluate the comparative statement. To create claims of this type, we mine Wikipedia by selecting a page *X* and extracting sentences with the pattern "is/was/named the *A* of *Y*" (e.g. *A* is *"first governor"*) where *Y* links to another page. Then we manually create temporal claims by examining dates on *X* and *Y* and describing the relation between the entities and events.

Named Entity Ambiguity and Lexical Variation On the FEVER task, a fact-checking system must resolve named entities and align words to make a prediction of veracity. We consider how subtle variations in entities and words may affect relation prediction.

ENTITY DISAMBIGUATION has been shown to be important for retrieving the correct page for an entity among multiple candidates [Hanselowski et al., 2018]. To create examples that contain ambiguous entities we selected claims from FEVER 1.0 development where at least one Wikipedia disambiguation page was returned by the Wikipedia python API.[12] We then created a new claim using one of the documents returned from the disambiguation list. We can create a query from a claim such as *"Patrick Stewart is someone who does acting for a living."* The query returns the page **[Patrick Stewart (disambiguation)]** which in turn gives a list of pages **[Patrick Stewart]**, **[Patrick Maxwell Stewart]**, and **[Patrick Stewart burial controversy]**. We can then create a new

[12] https://pypi.org/project/wikipedia/

claim which can be verified with evidence from the page **[Patrick Maxwell Stewart]**: *"Patrick Stewart was a London merchant."*

Finally, as previous work has shown that neural models are vulnerable to LEXICAL SUBSTITUTION [Alzantot et al., 2018], we apply their genetic algorithm approach to replace words and make a modified claim adversarial to a model fine-tuned on claims and their gold evidence sentences. We use counter-fitted word embeddings [Alzantot et al., 2018] to replace synonyms, hypernyms, or hyponyms, e.g. *created* → *established*, *leader* → *chief*. We then manually remove ungrammatical claims or incorrect relations.

Overall Dataset Our adversarial dataset contains 1000 examples, 417 of which are multi propositions, 313 are temporal and 270 are ambiguity and lexical variation. We present representative examples in Table 6.9, along with the corresponding label and evidence.

6.2.3 Methods

Verifying check-worthy claims such as those in Section 6.2.2 requires a system to 1) make sequential decisions to handle multiple propositions, 2) support temporal reasoning, and 3) handle ambiguity and complex lexical relations. To address the first requirement we make use of a pointer network [Vinyals et al., 2015] in two novel ways: i) to re-rank candidate documents and ii) to jointly predict a sequence of evidence sentences and veracity relations in order to compose evidence (Figure 6.4). To address the second we add a post-processing step for simple temporal reasoning. To address the third we use rich, contextualized representations. Specifically, we fine-tune BERT [Devlin et al., 2019] as this model has shown excellent performance on related tasks and was pre-trained on Wikipedia.

Our full pipeline is presented in Figure 6.3. We first identify an initial **candidate set of documents** (1a) by combining the top M pages from a TF-IDF search using DrQA [Chen et al., 2017a] with pages from the approach used in our prior work [Chakrabarty et al., 2018], which provides results from Google search and predicted named entities and noun phrases. Then, we perform

Attack Type	Example Claim	Label	Evidence
Conjunction	Blue Jasmine has Sally Hawkins acting in it **and** ~~Blue Jasmine~~ was filmed in San Francisco.	NEI	N/A
Multi-Hop Reasoning	Goosebumps was directed by ~~Rob Letterman~~ **the person who co-wrote Shark Tale**.	S	**[Goosebumps (film)]** It was directed by Rob Letterman, and written by Darren Lemke, based from a story by Scott Alexander and Larry Karaszewski. **[Rob Letterman]** Before Letterman's film subjects took him into outer space with Monsters vs. Aliens (2009), he was taken underwater, having co-directed and co-written Shark Tale.
Additional Unverifiable Propositions	Roswell is an American TV series **with 61 episodes**.	NEI	N/A
Date Manipulation	Artpop was Gaga's second consecutive number-one record in the United States ~~in 2009~~ **before 2010**.	R	**[Artpop]** Gaga began planning the project in 2011, shortly after the launch of her second studio album, Born This Way.
Multi-Hop Temporal Reasoning	Lisa Murkowski's father resigned from the Senate after serving as Senator.	S	**[Lisa Murkowski]** She is the daughter of former U.S. Senator and Governor of Alaska Frank Murkowski. Murkowski was appointed to the U.S. Senate by her father, Frank Murkowski, who resigned his seat in December 2002 to become the Governor of Alaska. **[Frank Murkowski]** He was a United States Senator from Alaska from 1981 until 2002 and the eighth Governor of Alaska from 2002 until 2006.
Entity Disambiguation	Kate Hudson is a left wing political activist	S	**[Kate Hudson (activist)]** Katharine Jane "Kate" Hudson (born 1958) is a British left wing political activist and academic who is the General Secretary of the Campaign for Nuclear Disarmament (CND) and National Secretary of Left Unity.
Lexical Substitution	The Last Song began ~~filming~~ **shooting** on Monday June 14th 2009.	R	**[The Last Song (film)]** Filming lasted from June 15 to August 18, 2009 with much of it occurring on the islandś beach and pier.

Table 6.9: Examples of the seven sub-types of attacks. Claims edited with word substitution or insertion have their changes in bold. Deletions are marked in strikethrough. Wikipedia titles are represented in bold with square brackets. **S:** SUPPORTS **R:** REFUTES **NEI:** NOT ENOUGH INFORMATION

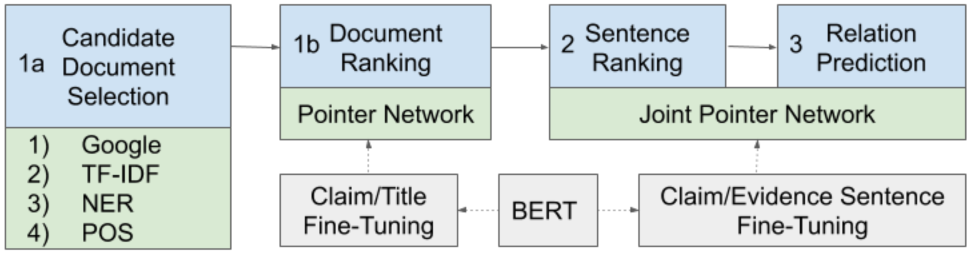


Figure 6.3: Our FEVER pipeline: 1) Retrieving Wikipedia pages by selecting an initial candidate set (1a) and ranking the top D (1b); 2) Identifying the top N sentences; 3) Predicting supports, refutes, or not enough info. Dashed arrows indicate fine-tuning steps.

document ranking by selecting the top $D < M$ pages with a pointer network (1b). Next, an N-long sequence of evidence sentences (2) and veracity relation labels (3) are **predicted jointly by another pointer network**.

Prior to training, we fine-tune BERT for document and sentence ranking on claim/title and claim/sentence pairs, respectively. The fine-tuned models give us representations that we can use as input to the pointer network. Each claim and evidence pair in the FEVER 1.0 dataset has both the title of the Wikipedia article and at least one sentence associated with the evidence, so we can train on these pairs directly. For every claim and title pair, we fine-tune the model to make a binary prediction of whether the evidence for the claim is likely to be found on the Wikipedia page associated with the title (instead of including the entire article text). For every claim and evidence sentence pair, we fine-tune the model to predict either SUPPORTS, REFUTES, or NOT ENOUGH INFO.

The core component of our approach is the pointer network, as seen in Figure 6.4. Given that the full evidence may consist of more than one document or sentence, we use the pointer network to re-rank candidate documents and jointly predict a sequence of evidence sentences and veracity labels. For the example in Figure 6.4, the claim *"Michelle Obama's husband was born in Kenya,"* we need the sequence of evidence sentences p_0 =*"Barack Obama was born in Hawaii"* and p_1 =*"Michelle married Barack Obama"* to identify that the claim is false. We thus predict the sequence of $N = 2$ sentences along with a sequence of $N = 2$ veracity labels for each timestep. In

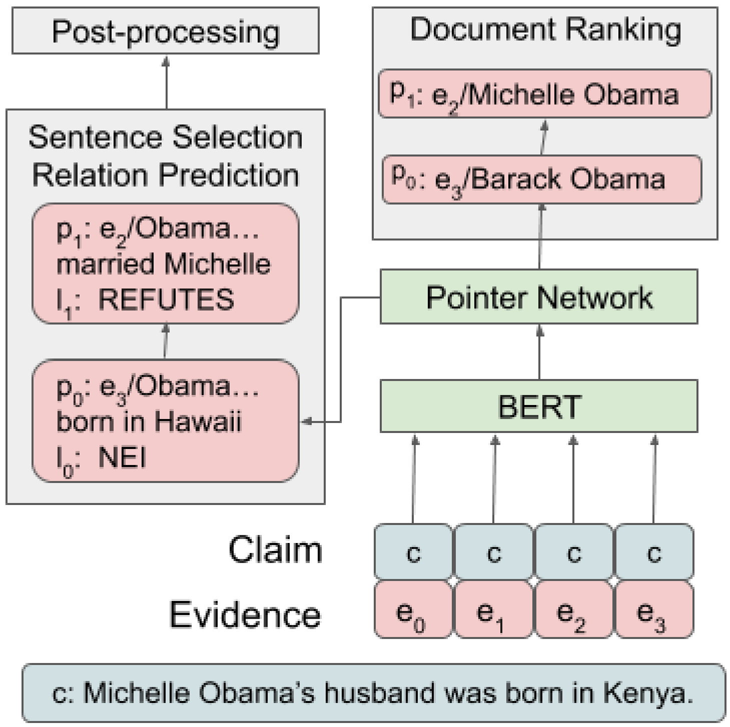


Figure 6.4: Pointer network architecture. Claim and evidence (page title or sentence) are embedded with BERT and evidence is sequentially predicted (for sentence selection the relation sequence is jointly predicted).

this example, at the first timestep, the evidence that Obama was born in Hawaii is not sufficient to label this claim as REFUTES so the first label in the sequence is l_0 = NEI. However, the cumulative evidence at the second timestamp is sufficient, as this links Michelle and Barack Obama via marriage, so the label at the second timestep is l_1 = REFUTES. For document ranking, the evidence sentences come from the Wikipedia articles with the titles **[Barack Obama]** and **[Michelle Obama]**, respectively, which we predict as a separate sequence of $D = 2$ titles.

Provided a candidate set of evidence E (as either document titles E_d where $|E_d| \geq D$ or sentences E_n where $|E_n| \geq N$) and a respective fine-tuned BERT model, we extract features for every claim c and evidence e_p pair by summing the $[CLS]$ embedding for the top 4 layers (as recommended by Devlin et al. [2019]):

$$m_p = BERT(c, e_p) \tag{6.7}$$

Next, to select the top k evidence, we use a pointer network over the evidence for claim c to extract evidence recurrently. The extraction probability[13] for evidence e_p at time $t < k$ is then:

$$u_t = \begin{cases} v_e^T \tanh(W[m_p; h_t^q]), & \text{if} p_t \neq p_s \forall s < t. \\ -\inf, & \text{otherwise.} \end{cases} \tag{6.8}$$

$$P(p_t | p_0 \cdots p_{t-1}) = \text{softmax}(u_t) \tag{6.9}$$

Then we compute the weighted average h_t^q[14] of the entire evidence set using q "hops" over the evidence [Vinyals et al., 2016, Sukhbaatar et al., 2015]:[15]

$$\begin{aligned} \alpha_t^o &= \text{softmax}(v_h^T \tanh(W_g m_p + W_a h_t^{o-1})) \\ h_t^o &= \sum_j \alpha_t^o W_g m_j \end{aligned} \tag{6.10}$$

We concatenate m_p and h_t^q and use a multi-layer perceptron (MLP) to predict p_t. The loss is then:

$$\mathcal{L}(\theta_{ptr}) = -1/k \sum_{t=0}^{k-1} \log P_{\theta_{ptr}}(p_t | p_{0:t-1}) \tag{6.11}$$

We train on gold evidence and perform inference with beam search for both **document ranking** and **joint sentence selection and relation prediction**.

Document Ranking In order to obtain representations as input to the pointer network for document ranking, we leverage the fact that Wikipedia articles all have a title (e.g. **[Barack Obama]**), and fine-tune BERT on title and claim pairs, in lieu of examining the entire document text (which due to its length is not suitable for BERT). Because the title often overlaps lexically with the claim (e.g. **[Michelle Obama]**), we can train the model to locate the title in the claim. Furthermore, the words in the title co-occur with words in the article (e.g. *"Barack"* and *"Michelle"*), which the

[13] Set to $-inf$ only while testing

[14] Initially, $h^{t,0}$ is set to z_t, the hidden state of the pointer network decoder LSTM.

[15] v_h, W_g, and W_a are learned parameters.

pre-trained BERT language model may be attuned to.

We thus create a dataset for fine-tuning by extracting all titles of gold evidence Wikipedia pages from the FEVER 1.0 training set. However, this provides us with only positive training examples, so to obtain negative samples, we randomly sample titles from pages retrieved during our Candidate Document Selection stage. In other words, for each positive pair (gold title and claim), we sample title pages to obtain irrelevant ones. This results in 140,085 positive examples and 630,265 negative examples for training with approximately 10% set aside for validation (16,016 positive examples and 84,437 negative). We then fine-tune a BERT model on this dataset, obtaining 90.0% accuracy. Figure 6.5 depicts an example training datapoint and the BERT representation. Given the fine-tuned model, we then extract features using Equation 6.7 where e_p is a title, and use Equation 6.11 to learn to predict a sequence of titles as in Figure 6.4.

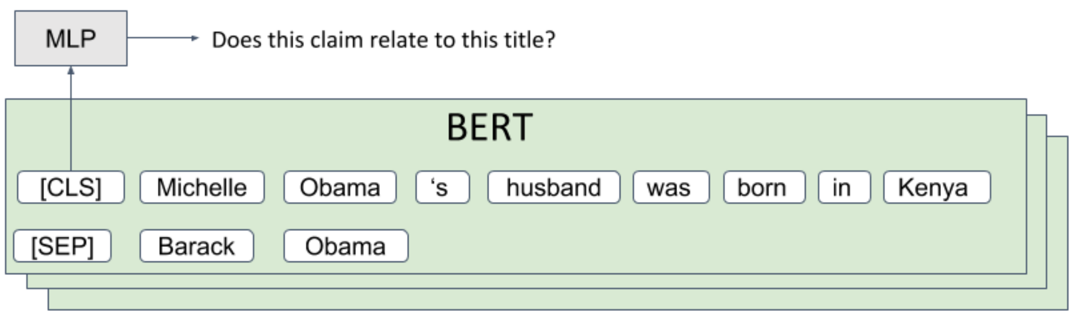


Figure 6.5: An example claim/title pair training instance.

Joint Sentence Selection and Relation Prediction The sentence selection and relation prediction tasks are closely linked, as predicting the correct evidence is necessary for predicting SUPPORTS or REFUTES and the representation should reflect the interaction between a claim and an evidence set. Conversely, if a claim and an evidence set are unrelated, the model should predict NEI. We thus jointly model this interaction by sharing the parameters of the pointer network - the hidden state of the decoder is used for both tasks and the models differ only by a final MLP.

For **sentence selection**, similar to our document selection fine-tuning approach, we fine-tune a classifier on claim and evidence sentence pairs to obtain BERT embeddings. However, instead of training a binary classifier for the presence of valid evidence, we train directly on veracity prediction, which is better suited for the end task. We use the FEVER 1.0 training set to obtain the gold veracity relation labels for each claim (SUPPORTS, REFUTES, and NEI). For SUPPORTS and REFUTES claims, the training set also has gold evidence sentences, which we use as provided. As gold evidence is not available for NEI relations, we sample sentences from our retrieved candidate documents to maintain a balanced dataset. We also enhance each evidence sentence by *prepending the Wikipedia article title*, as many sentences from Wikipedia require co-reference resolution and we make the assumption that the title of the article can resolve the co-reference. An example is given in Figure 6.6. Our dataset of sentence and claim pairs consists of 54,431 SUPPORTS relations, 54,592 REFUTES relations, and 54,501 NEI relations in training, with approximately 10% set aside for validation (6,139 SUPPORTS relations, 5,984 REFUTES relations, and 6,050 NEI relations). We then fine-tune a BERT classifier on relation prediction, obtaining 93% accuracy. Given the fine-tuned model, we extract features using Equation 6.7 where e_p is a sentence, and use Equation 6.11 to learn to predict a sequence of sentences.

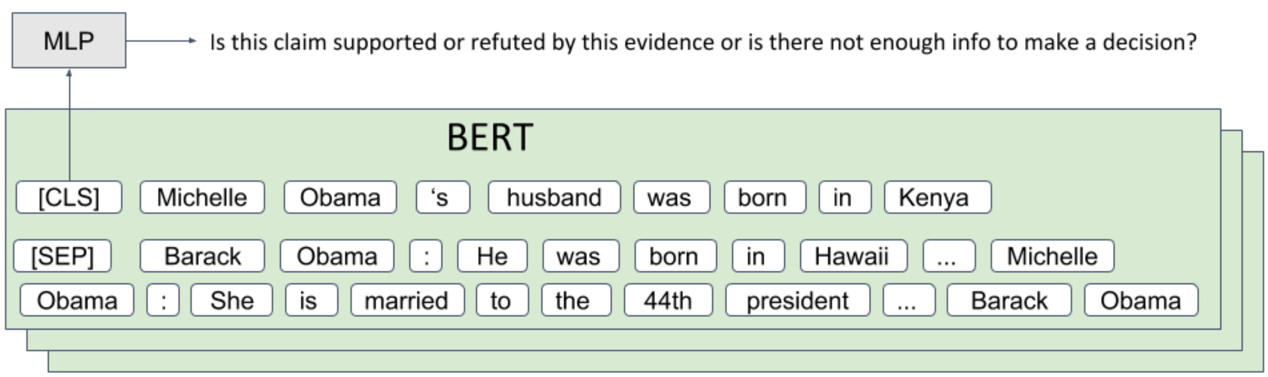


Figure 6.6: An example claim/evidence pair training instance.

In order to closely link **veracity relation prediction** with evidence prediction, we re-frame the task as a sequence labeling task. In other words, rather than make a single prediction given all evidence sentences, we make one prediction at every timestep during decoding to model the relation between the claim and *all evidence retrieved to that point*. This approach provides three benefits: it allows the model to better handle noise (when an incorrect evidence sentence is predicted), to

Claim: Michelle Obama's husband was born in Kenya.
Evidence: [Barack Obama] Obama was born in Honolulu, Hawaii.
Evidence: [Michelle Obama] She is married to the 44th President of the United States Barack Obama.
Label: REFUTES

Figure 6.7: Multi-Hop Reasoning Example

Claim: Murda Beatz′s real name is Marshall Mathers.
Evidence: [Murda Beatz] Shane Lee Lindstrom (born February 11, 1994), known professionally as Murda Beatz, is a Canadian hip hop record producer and songwriter from Fort Erie, Ontario.
Label: REFUTES

Figure 6.8: Single-Hop Reasoning Example

handle multi-hop inference (to model the occurrence of switching from NEI to S/R), and to effectively provide more training data (for $k = 5$ timesteps we have 5 times as many relation labels). For the claim in Figure 6.4 (full text provided in Figure 6.7), the initial label sequence is NEI and REFUTES because the first evidence sentence by itself (the fact that Barack Obama was born in Hawaii) would not refute the claim. Furthermore for $k = 5$, the remaining sequence would be REFUTES, REFUTES, REFUTES, as no matter what additional evidence is predicted, the prediction of REFUTES would not change, as evidence is guaranteed to be non-contradictory in FEVER. On the other hand, given a claim that requires only a single piece of evidence, such as that in Figure 6.8, the sequence of relation labels would be REFUTES, REFUTES, REFUTES, REFUTES, REFUTES if the correct evidence sentence was selected at the first timestep, NEI, REFUTES, REFUTES, REFUTES, REFUTES if the correct evidence sentence was selected at the second timestep, and so forth.

We augment the evidence sentence selection described previously to use the hidden state of the pointer network after q hops (Equation 6.10) and an MLP to also predict a veracity relation label at that time step, closely linking evidence and relation prediction:

$$P(l_t) = \text{softmax}(W_{l2}\tanh(W_{l1}h_t^o)) \tag{6.12}$$

As with evidence prediction (Equation 6.11), when the gold label sequence is available, the loss term is:

$$\mathcal{L}(\theta_{rel\_seq}) = -1/k \sum_{t=0}^{k-1} \log P_{\theta_{rel\_seq}}(l_t) \tag{6.13}$$

When training, at the current timestep we use both the gold evidence, i.e. "teacher forcing" [Williams and Zipser, 1989], and the model prediction from the previous step, so that we have training data for NEI. Combining Equations 6.11 and 6.13, our loss is:

$$\mathcal{L}(\theta) = \lambda \mathcal{L}(\theta_{ptr}) + \mathcal{L}(\theta_{rel\_seq}) \tag{6.14}$$

Finally, to predict a relation at inference, we ensemble the sequence of predicted labels by averaging the probabilities over every time step.[16]

Post-processing for Simple Temporal Reasoning Because neural models are unreliable for handling numerical statements, we introduce a post-processing rule-based step to extract and reason about dates. We use the Open Information Extraction system of Stanovsky et al. [2018] to extract tuples of natural language expressions representing basic propositions asserted by the claim or the evidence. For example, given the claim *"The Latvian Soviet Socialist Republic was a republic of the Soviet Union 3 years after 2009."* the system would identify the verb *"was"* and the text preceding and following the verb as **ARG0** and **ARG1**. After identifying tuples in both claims and predicted evidence sentences, we discard those lacking dates (e.g. discarding **ARG0**). Given more than one candidate evidence sentence, we select the one with either maximum lexical overlap or selected earlier by the pointer network.[17]

Once we have both the claim date-tuple and evidence date-tuple we apply the following rules:

1. If there is a date range (e.g. *between/during/in*), evaluate whether the evidence date is between two dates in the claim

[16]The subset of timesteps was determined empirically: while at the final timestep the model is likely to have seen the correct evidence it also contains more noise; in future work we will experiment with alternatives.

[17]This process introduces additional uncertainty, as we do not know if the dates in the evidence align with the dates in the claim. Future work could introduce a soft alignment via latent variables or an attention mechanism.

2. If there is an offset (e.g. *x years/days before/after*), add/subtract x years from the date in the claim and compare to the evidence date

3. If there is a temporal expression (e.g. *before/after/in*, compare the claim date directly to the evidence date

For instance, for the date expression *"3 years after 2009,"* the algorithm compares the year *"2012"* to the date in the retrieved evidence (*"1991,"* the year the USSR dissolved) and labels the claim as REFUTES.

6.2.4 Experiments and Results

We evaluate our dataset and system as part of the FEVER 2.0 shared task. The main goal of the evaluation was to test the vulnerabilities introduced by our adversarial claims (Section 6.2.2) and how well our system (Section 6.2.3) addressed those vulnerabilities. We train on the FEVER 1.0 training data described in Section 6.2.1 and evaluate our system on the FEVER 1.0 and 2.0 development and blind test sets.

Performance Measures We use a number of measures to evaluate our dataset and system. *Accuracy* refers to the percentage of labels predicted correctly. *Recall* is defined as whether the gold evidence is contained in selected evidence at $k = 5$. The *FEVER score* is then the percentage of correct labels that also have correct evidence (or are correctly predicted in the NEI case). Finally, *Potency* is the inverse FEVER score (subtracted from one) of a dataset macro-averaged over multiple systems and is used to evaluate adversarial claims.

Our Baseline-RL For baseline experiments, to compare different loss functions, we use the approach of Chakrabarty et al. [2018] for document selection and ranking, the reinforcement learning (RL) method of Chen and Bansal [2018] for sentence selection, and BERT [Devlin et al., 2019] for relation prediction. The RL approach using a pointer network is detailed by Chen and Bansal [2018] for extractive summarization, with the only difference that we use our fine-tuned BERT on claim/gold sentence pairs to represent each evidence sentence in the pointer network (as with our full system) and use the FEVER score as a reward. The reward is obtained by selecting sentences with the pointer network and then predicting the relation using an MLP (updated during training) and the concatenation of all claim/predicted sentence representations with their maximum/minimum pooling. We use RL to compare the use of different loss functions and we directly maximize the value we are interested in: the FEVER score; this encourages the model to focus on extracting the sentences that allow it to correctly predict the relation label. To make the sentence extractor an RL agent, we can formulate a Markov Decision Process (MDP): at each extraction step t, given a claim c, the agent observes the current state and samples an action from Equation 2 to extract a document sentence s, predict the relation label l and receive a reward $r(t+1) = \text{FEVER}(c, s, l)$. We train using REINFORCE, adapted with an Actor-Critic to minimize variance (detailed by Chen and Bansal [2018]). As RL often requires pre-training, we combine the pointer network loss from Equation 3 with the RL loss ($\mathcal{L}(\theta_{rl})$) and the relation prediction loss ($\mathcal{L}(\theta_{rel})$:

$$\mathcal{L}(\theta) = \lambda_1 \mathcal{L}(\theta_{ptr}) + \lambda_2 \mathcal{L}(\theta_{rl}) + \mathcal{L}(\theta_{rel}) \tag{6.15}$$

We set both $\lambda_1 = 1$ and $\lambda_2 = 1$.

Hyper-parameters and Experimental Settings We select $M = 30$ Wikipedia articles using TF-IDF as part of our candidate document selection step and select $D = 5$ after document ranking. We select $N = 5$ sentences during sentence selection, consistent with the shared task evaluation. We use version 0.5.0 of the Huggingface library[18] to fine-tune the "BERT-base" model on claim/title and claim/evidence pairs using the default settings. We lowercase all tokens and use the default BERT tokenizer. As recommended by Devlin et al. [2019], we select hyper-parameters by grid search on the validation set over 16 and 32 for batch size, 2e-5, 3e-5, and 5e-5 for learning rate, and 3 and 4 for the number of epochs. Given the respective fine-tuned models, we train both the document ranking and sentence selection pointer networks on the FEVER 1.0 training sets using

[18] `https://github.com/huggingface/pytorch-pretrained-BERT`

Adagrad [Duchi et al., 2011] with a learning rate of 0.01, a batch size of 16, and a maximum of 140 epochs with early stopping on the FEVER 1.0 development set. The dimension of the pointer network LSTM hidden state is set to 200 with $q = 3$ hops over the memory. We use a beam width of 5 during inference. The MLP used to predict relations has a hidden layer dimensionality of 200 and we set $\lambda = 1$.

Adversarial Dataset Evaluation We present the performance of our adversarial claims, obtained by submitting to the shared task server, which reports the performance of several top-performing systems on this set of claims. We compare our claims to other participants in the FEVER 2.0 shared task (Table 6.10) and divided by attack type (Table 6.11). Table 6.10 reports the raw potency (inverse FEVER score macro-averaged across the top-performing fact-checking systems [Thorne and Vlachos, 2019]), correctness (percentage of claims with correct evidence and veracity as annotated by the organizers) and adjusted potency (potency of correct examples); these measures were all calculated by the organizers.

Compared to other participants (Table 6.10), we presented a larger set of claims (the organizers used 501 of our claims for development and 499 for test). We rank second in adjusted potency, but we provided a more diverse set than those created by the organizers or other participants. The organizers [Thorne and Vlachos, 2019] created adversarial claims using simple pattern-matching and replacement, e.g. quantifiers and negation. Niewinski et al. [2019] trained a GPT-2-based model on the FEVER data and manually filtered disfluent claims. Kim and Allan [2019] considered a variety of approaches, the majority of which required understanding area comparisons between difference regions (e.g. that Nerva, Span is larger than Madhya Pradesh) or understanding implications (e.g. that "not clear" implies NEI). While GPT-2 is effective, our approach is controllable and targeted at real-world challenges. Finally, Table 6.11 shows that when we select our top 200 most effective examples (multi-hop reasoning and multi-hop temporal reasoning) and compare to the approaches of Niewinski et al. [2019] and Kim and Allan [2019] (who both provided less than 204 examples total) our potency is much higher. In particular, multi-hop reasoning has a potency of 88% for SUPPORT relations and 93% for REFUTES relations, which is at least 21% higher than the approach of Niewinski et al. [2019]. Additionally, multi-hop temporal reasoning is extremely potent, with a 98% potency for both SUPPORT and REFUTES relations.

Team	#	Raw	Correct	Potency
Baseline - Thorne and Vlachos [2019]	498	60.3%	82.33%	49.6%
Kim and Allan [2019]	102	79.7%	64.71%	51.5%
Ours	**501**	**68.5%**	**81.44%**	**55.8%**
Niewinski et al. [2019]	79	79.9%	84.81%	66.83%

Table 6.10: The evaluation of our claims in relation to the other participants **#:** Number of examples in blind test **Raw:** Raw potency score **Correct:** Percent grammatical and coherent with correct label and evidence **Potency:** Adjusted potency (potency of correct examples)

Attack	M/A	#S	%S	#R	%R	#NEI	%NEI
CONJUNCTION	A	-	-	54	55%	75	63%
MULTI-HOP REASONING	M	100	88%	88	93%	99	50%
ADDITIONAL UNVERIFIABLE PROPOSITIONS	M	-	-	-	-	50	50%
DATE MANIPULATION	A	49	59%	129	80%	80	46%
MULTI-HOP TEMPORAL REASONING	M	46	98%	5	98%	4	29%
ENTITY DISAMBIGUATION	M	46	50%	-	-	-	-
LEXICAL SUBSTITUTION	A*	92	70%	57	70%	25	38%

Table 6.11: **Attack:** Type of attack as described in Section 6.2.2 **M/A:** Whether the claims are generated manually (M) or automatically (A) or verified manually (A*) **#S:** Number of support examples **#R:** Number of refute examples **#NEI** Number of not enough info examples **P:** Potency on Shared Task systems.

Evaluation against State-of-the-art In Table 6.12 we compare Our System (Section 6.2.3) to recent work from teams that submitted to the shared task server for FEVER 1.0 and 2.0, including the results of Our Baseline-RL system. For the FEVER 1.0 shared task, all measures (FEVER score, accuracy and evidence recall at $k = 5$) were reported but the organizers only reported the FEVER score for the FEVER 2.0 shared task. If a system was not submitted to the shared task server, the corresponding column is indicated with a dash. The table includes the shared task baseline [Thorne et al., 2018, Thorne and Vlachos, 2019] and the top four systems from FEVER

1.0 [Hanselowski et al., 2018, Malon, 2018, Yoneda et al., 2018, Nie et al., 2018].

Our dual pointer network approach obtains state-of-the-art results on the FEVER 1.0 blind test set (Table 6.12) on all measures even over systems designed specifically for evidence retrieval [Nishida et al., 2019a, Zhou et al., 2019], largely due to a notable improvement in recall (more than 3 points over the next system [Hanselowski et al., 2018]). We also find improvements in accuracy over the remaining pipeline systems, suggesting that joint learning helps. Compared to Our Baseline-RL, Our System has 1.8 point improvement in FEVER score on FV1-test with 4 points on FV2-test. Notably, our system finishes second (with a score of 36.61) on the FEVER 2.0 shared task test set, although our claims were designed to be especially challenging with regards to our model. The model of Malon [2018] performs especially well; they also use a transformer-based architecture although without a pre-trained model such as BERT.

	FEVER 1.0 blind test			FEVER 2.0 blind test
System	Acc.	Rec.	FEVER score	FEVER score
Thorne et al. [2018]	48.84	45.89	27.05	11.06
Malon [2018]	50.02	61.08	57.36	**37.31**
Hanselowski et al. [2018]	65.46	85.19	61.58	25.35
Nishida et al. [2019a]	69.30	76.30	61.80	-
Yoneda et al. [2018]	67.62	82.84	62.52	35.83
Nie et al. [2018]	68.16	71.51	64.21	30.47
Our Baseline-RL	-	-	67.08	32.92
Zhou et al. [2019]	71.60	-	67.10	-
Our System	**72.47**	**88.39**	**68.80**	36.61

Table 6.12: Comparison with state of the art on FEVER 1.0 and 2.0 blind test

System Component Ablation To better understand the improved performance of our system, we present two ablation studies. Table 6.13 presents the effect of using different objective functions for sentence selection and relation prediction, compared to the loss term used in our full model from Equation 6.14. Table 6.14 evaluates the impact of the document pointer network and rule-based date handling on our adversarial claims. Our system is significantly better on all measures ($p < 0.001$ by the approximate randomization test).

Table 6.13 presents the effect of using different objective functions for sentence selection and relation prediction, compared to joint sentence selection and relation prediction in our full model. We compare Our System to Our Baseline-RL system as well as another baseline (*Ptr*). The *Ptr* system is the same as Our Baseline-RL, except the pointer network and MLP are not jointly trained with RL but independently using gold evidence and predicted evidence and relations, respectively. Finally, the Oracle upper bound presents the maximum possible recall after our document ranking stage, compared to 94.4% for Chakrabarty et al. [2018], and relation accuracy (given the MLP trained on 5 sentences guaranteed to contain gold evidence). We find that by incorporating the relation sequence loss, we improve the evidence recall significantly relative to the oracle upper-bound, reducing the relative error by 50% while also obtaining improvements on relation prediction, even over a strong RL baseline. Overall, the best model is able to retrieve 95.9% of the possible gold sentences after the document selection stage, suggesting that further improvements are more likely to come from document selection.

Model	Acc.	Rec.	FEVER
Oracle	84.2	94.7	–
Ptr	74.6	86.1	68.6
Our Baseline-RL	74.6	87.5	69.2
Our System	76.74	90.84	73.17

Table 6.13: Ablation experiments on FEVER 1.0 development set

Table 6.14 evaluates the impact of the document pointer network and rule-based date handling on FV2-dev, as the impact of multi-hop reasoning and temporal relations is less visible on FV1-dev. We again compare Our Baseline-RL system to Our System and find an even larger 7.16 point improvement in FEVER score. We find that ablating the date post-processing (*-dateProc*) and both the date post-processing and document ranking components (*-dateProc,-docRank*) reduces the FEVER score by 1.45 and 3.5 points, respectively, with the latter largely resulting from a 5 point decrease in recall.

Ablation for Claim Types While Table 6.11 presents the macro-average of all systems by attack type, we compare the performance of Our Baseline-RL and Our System on different types of

System	Acc.	Rec.	FEVER
Our System	48.13	63.28	43.36
-dateProc	45.14	63.28	41.91
-dateProc,-docRank	44.29	58.32	39.86
Our Baseline-RL	44.04	57.56	36.2

Table 6.14: Ablation experiments on FEVER 2.0 development set

attacks in Table 6.15.

Our System improves on evidence recall for multi-hop reasoning and multi-hop temporal reasoning types (indicating that a multi-hop document retrieval step may help) and those involving entity disambiguation or lexical substitution (using a model to re-rank may remove false matches with high lexical similarity). For example, the claim *"Honeymoon is a major-label record by Elizabeth Woolridge Grant."* requires multi-hop reasoning over entities. Our System correctly retrieves the pages **[Lana Del Rey]** and **[Honeymoon (Lana Del Rey album)]**, but Our Baseline-RL is misled by the incorrect page **[Honeymoon]**. While Our System outperforms Our Baseline-RL in terms of recall, one area for improvement is on multi-hop claims in terms of accuracy, which decreases even though recall increases. In the baseline system, accuracy is higher than recall. This should not occur if the model is learning a relation between a claim and an evidence, suggesting that it may be learning a bias of the label distribution or the claim alone.

We also obtain large improvements on date manipulation examples (here a rule-based approach is better than our neural one); in contrast, multi-hop temporal reasoning leaves significant room for improvement. For instance, for the claim *"The MVP of the 1976 Canada Cup tournament was born before the tournament was first held,"* our full system correctly retrieves **[Bobby Orr]** and **[1976 Canada Cup]** (unlike the RL baseline). However, a further inference step is needed beyond our current capabilities – reasoning that Orr's birth year (1948) is before the first year of the tournament (1976).

Finally, we enhance performance on multi-propositions as conjunctions or additional unverifiable information (indicating that relation sequence prediction helps). Claims (non-verifiable phrase in brackets) such as *"Taran Killam is a [stage] actor."* and *"Home for the Holidays stars an ac-*

Attack Type		Accuracy	Recall	FEVER
Conjunction	B	16.95	92.0	16.95
	S	**40.68****	92.0	**40.68****
Multi-hop Reasoning	B	55.81*	29.07	19.77
	S	33.72	**45.35***	17.44
Additional Unverifiable Propositions	B	48.0	–	48.0
	S	**80.0****	–	**80.0****
Date Manipulation	B	30.99	79.59	27.46
	S	**53.52*****	79.59	**42.25****
Multi-hop Temporal Reasoning	B	3.33	10.34	0.0
	S	3.33	**13.79**	0.0
Entity Disambiguation	B	70.83	62.5	58.33
	S	**79.17**	**79.17***	**70.83**
Lexical Substitution	B	33.33	65.71	25.0
	S	29.76	**75.71***	**26.19**

Table 6.15: Attack results for our FV2-dev claims. **B:** Our Baseline-RL, **S:** Our System. *: $p < 0.05$ **: $p < 0.01$ ***: $p < 0.001$ by approximate randomization test

tress [born in Georgia].” are incorrectly predicted by the baseline even though correct evidence is retrieved.

6.2.5 Conclusions, Limitations, and Future Work

We demonstrated current weaknesses in approaches to fact-checking and created a novel dataset of adversarial claims to evaluate new developments. We took steps towards realistic fact-checking by implementing targeted improvements to multi-hop reasoning (using a document pointer network and a pointer network for sequential joint sentence selection and relation prediction), simple temporal reasoning (by rule-based date handling), and ambiguity and lexical variation (by fine-tuning contextualized distributed representations).

We hope that the release of our dataset will lead to improved state of the art particularly for multi-hop and temporal reasoning. For multi-hop reasoning, one possibility is to explore modeling the relation between the claim and the full evidence sequence, rather than sequential claim and evidence pairs. While we model the relation between claim and evidence using the BERT pair representation, one limitation of the pointer network approach is that we do not directly model the relationship between the claim and the full set of evidence selected at a particular timestep (e.g. by concatenating the evidence sentences). Future work could address this knowledge gap in the model. Future work could also build on work in multi-hop question answering [Nishida et al., 2019b] and continue to examine the link between question answering and fact checking. For temporal reasoning, our rule-based query reformulation step improves over a fine-tuned BERT model but requires manual management of a number of different cases. Future work could thus investigate techniques for incorporating numerical reasoning [Andor et al., 2019]. For multi-hop temporal reasoning, this type of reasoning should be incorporated into the sentence selection stage to improve recall as well.

Another area of improvement is in reference disambiguation. For the example provided in Section 6.2.3, we assume that *"Michelle Obama's husband"* refers to Barack Obama but this would not be the case if she had been married more than once. In general, in the multi-hop inference scenario, making an assumption that results in a reference to the wrong entity would cause the entire chain to be incorrect. This example raises the question of what the correct disambiguation should be; there are multiple people that share identifiers such as a name (e.g. Steve McQueen) and one possibility is to disambiguate using some proxy for prominence. Recent work has found that language models tend to disambiguate using only a first name, which tends to correspond to their notability (e.g. Donald for Donald Trump) [Shwartz et al., 2020]. However, regardless of the fame of a given entity, accurate modeling likely requires providing more context than a single sentence in the case of FEVER. In a real-world scenario, check-worthy claims occur in social media or in news articles and we thus have access to an entire post (or thread) or column. Thus, making progress towards realistic fact-checking also requires modeling claims that would occur in these settings and the full context in which they occur.

Another limitation is that our model is focused on recall rather than precision. As the FEVER metric allows up to 5 evidence sentences, even though the majority of cases only require one evidence sentence, precision at 5 is low. Although the model is fairly accurate at reasoning over noisy evidence, in order to provide retrieved evidence as part of an argument, we need higher-precision models. One possibility is to incorporate kernel methods [Tymoshenko et al., 2017] into the evidence selection stage, as these models have demonstrated excellent performance on answer sentence re-ranking for question answering.

Finally, there are limitations of both the FEVER 1.0 and 2.0 datasets. Due to the nature of Wikipedia, which focuses on entities and concepts, the claims from FEVER 1.0 which were generated based on these articles tend to be *attribute-based* and the true properties of these entities can be found in corresponding Wikipedia articles. However, many realistic claims tend to be *event-based*, which involve multiple entities and concepts. Our adversarial dataset addresses some of these issues by providing multi-hop and temporal reasoning claims, but there are many other types of realistic challenges for fact-checking. One specific challenge is that of fact-checking for causal relations, as discussed in Section 6.1. While causal relations can be found in Wikipedia, according to our work in Section 6.1.2, detecting a causal relation is a different prospect than verifying a causal relation, as Wikipedia often provides multiple viewpoints on controversial topics. For example, on the topic of violence in video games, Wikipedia has the sentence *"A common theory is that playing violent video games* ***increases*** *aggression in young people."* A naive approach might fact check this statement as true, but any approach should account for the fact that this statement is reported belief [Diab et al., 2009], and label the statement as true *according to the proponent*. In regards to another challenge, according to the types of verifiable propositions described by [Park and Cardie, 2014], the claims in the FEVER dataset are verifiable non-experiential, but verifiable experiential claims (i.e. personal testimonies) provide a separate set of challenges. The FEVER claims often contain the Wikipedia article as the subject of the sentence, whereas verifying a personal testimony would require predicting whether it is possible for an event to have happened, a much more difficult problem. Lastly, accounting for style and domain is another difficult problem. A query reformulation step, or re-writing the claim to be similar to the FEVER claims, may help with this problem.

6.3 Application - Household Electricity Consumption

In Sections 6.1 and 6.2 we introduced methods and datasets for predicting causality and veracity. In order to demonstrate the importance of causal relations and fact verification in argumentation, we develop a system that mines claims from social media and presents them to experiment participants with the goal of positive behavior change. In Section 4.3.3, we showed that providing new content is one key aspect to changing someone's view. Similarly, our hypothesis is that providing novel information to recipients can affect changes in their behavior. We make a similar assumption that just as debaters act in good faith, people want to make positive changes but lack the knowledge to do so. In order to measure this hypothesis, we first need to obtain novel information. As we need a high-precision approach, we use our model of causal relations combined with a set of topic-related keywords to mine claims from social media. Second, we need to make sure that we are providing factual information so that the recipients can take the correct actions to modify their behaviors. Finally, we need an experimental setting where we can measure behavior change in a controlled environment. We thus conduct our experiments in the domain of household electricity consumption, where we are able to measure the daily consumption of participants and how it changes over time.

As part of a larger experiment, we have access to the regular electricity consumption of apartments in Columbia housing. The apartment-dwellers have the option to sign up to receive regular feedback on their energy usage. As part of this feedback, we send our mined suggestions for reducing consumption. The claims we mine would ideally appear as one of the following examples, with the causal relation AltLex highlighted in boldface:

1. Let your dishes air dry; if you don't have an automatic air-dry switch, turn off the control knob after the final rinse and prop the door open slightly **so** the dishes will dry faster.

2. Maintain your refrigerator at 37 to 40 degrees for efficiency, and freezer at 5 degrees – this will **save** you money and better serve you!

3. Energy efficient windows will help you **reduce** your carbon footprint and your energy bills.

While these examples all provide accurate advice for how to reduce electricity usage, small changes in these claims as they are passed around on social media could result in inaccurate information. The domain knowledge required to accurately label these claims for veracity is extensive. Therefore, we fact check the mined tips manually using expert and non-expert annotators.

We first mine claims from social media (discussed in Section 6.3.1) by using our methods for causal relation extraction in Section 6.1.3. Next, the claims are verified by expert annotators from electrical and mechanical engineers working in the field of household energy consumption (Section 6.3.1) and we compare the gold annotated data from the experts to that of non-experts with access to Google search. Then, although our primary focus is on the behavioral impact of our mined tips, we evaluate the performance of our system from Section 6.2.3 at predicting the veracity of these claims, along with other state-of-the-art methods (Section 6.3.2). The claims are subsequently presented as part of a system that uses sensors to monitor household electricity consumption and sends e-mails to voluntary participants (Section 6.3.2). Finally, we show that providing information to participants about how to reduce their electricity consumption reduces their usage compared to those participants that received no information (Section 6.3.3).

6.3.1 Data

Our experimental goals are two-fold: 1) verify our hypothesis that novel information can assist with positive behavior change and 2) to validate the importance of causal relations and fact verification. In order to address the first goal, we need a way to obtain novel information, which also allows us to address the second goal. In order to address the second goal, we automatically collect a dataset using a causal relation classifier to mine energy-related claims from social media and manually evaluate these claims for their veracity using expert and non-expert annotators.

Causal Relations We select a number of subreddits[19] and hashtags[20] that are likely to include suggestions for saving energy. For Twitter, we use the hashtags `#savingenergy`, `#energysaving`,

[19]from `https://reddit.com`
[20]from `https://twitter.com`

Category/Appliance	Keywords
Phantom Load	vampire, phantom, standby, leaking
Air Conditioner	conditioner, a/c, fan
Window	window
Space/Water Heater	heater
Fridge	fridge, freezer, refrigerator
Washing Machine	washing, washer
Dishwasher	dishwasher, washer, dish
Toaster	toaster
Computer	computer
Kettle	kettle
Television	tv, television
Dryer	dryer
Microwave	microwave

Table 6.16: The keywords used to identify sentences related to electricity consumption and their corresponding category.

`#energytips`, and `#energyefficiency`. For Reddit, we use the subreddits `/r/efficiency`, `/r/energyefficiency`, and `/r/frugal`. We further refine our search to require that posts/tweets contain one of several pre-defined keywords, under the assumption that an energy-saving tip must contain the appliance or aspect of energy-saving. We selected keywords according to a number of different categories, including those related to windows, dishwashers, or the "phantom load,"[21] i.e. the electricity consumed by appliances that are plugged in but not in use. The full set of keywords, along with the corresponding category are included in Table 6.16. We collect data between the dates of 2018/02/15 and 2018/04/26, resulting in 5,279 Tweets and 9,615 Reddit posts.

Next, to identify tips which provide methods to reduce electricity consumption, we first sentence and word tokenize each post with Spacy[22] before applying our causal relation classifier described in Section 6.1.3. For each sentence in the post, we consider all possible AltLexes, given the derived set of causal AltLexes after the bootstrapping stage described in Section 6.1.3. Then for each candidate AltLex, we predict whether a causal relation exists. If our classifier predicts a causal relation anywhere in the sentence, we add the sentence, along with the left and right context, to our set of claims. After this stage, we have 2,661 sentences from Twitter and 3,789 from Reddit.

[21] `https://en.wikipedia.org/wiki/Standbypower`
[22] `https://spacy.io`

Fact Verification After our data collection step, we obtain 6,450 total claims. However, while our method of data collection using causal relations provides us with data using a specific argument scheme, claims extracted from social media may be unreliable due to inaccurate information or noise from the extraction process. We assume that all claims identified in this process are verifiable (either experiential or factual) and we thus only need to verify claims instead of requiring the additional step of check-worthiness prediction.

Next, we conducted two rounds of annotations, with expert annotators and non-experts. We first recruited three expert annotators with backgrounds in mechanical and electrical engineering and experience in the field of household electricity consumption. All annotators were part of Columbia University - two professors and one graduate student. During annotation, we presented the selected claims to the annotators along with the one-sentence left context and right context and the keywords. We used an annotation scheme inspired by previous work on fact-checking in community forums [Mihaylova et al., 2018], which included the categories of true, false, partially true, conditionally true, responder unsure and non-factual. In our work, we use most of the categories except for "responder unsure," as this category was specific to the question-answer nature of their dataset. We also consider "irrelevant" instead of "non-factual," as our method for obtaining candidate claims may retrieve claims that do not provide suggestions for how to save electricity. The annotators were provided with the following definitions:

- **true:** if a recipient were to follow the advice given in the claim, it would reduce their electricity consumption. Some examples include *"Filters for air conditioners should be cleaned regularly"* and *"Don't let frost build up in the freezer compartment as this increases energy consumption."*

- **false:** if a recipient were to follow the advice given in the claim, it would not reduce their electricity consumption. Some examples known to be false are the claims *"Think about the following points when buying a new washing machine: Front-loaders are usually more efficient than top-loaders and use less detergent."* and *"for space heating that won't trip the breakers you can use any electrical device, the watts to btus is almost exactly constant* *across nearly all electrical devices so an old computer or a tv or something will heat up a room pretty comfortably if it's well insulated."*

- **irrelevant:** the claim is not an energy-saving suggestion. In our candidate set of claims, these are examples where they keywords are mentioned but the sentence may be off-topic, such as *"I know we have some habits that contribute to the increased cost, such as an entertainment system which is almost always on (HTPC, AV receiver and 55" LED TV), and constant laundry due to having 2 young kids, 1 of which is still in cloth diapers, full-time."*

- **partially true:** not all of the propositions can be verified or some references are not clear. For instance, in both of the following claims, it is not clear how to resolve the co-reference without additional context. In the claim, *"It's safer and faster than those slow, metal, oven-top kettles that I've always been annoyed using."* it is not clear what *"it"* is referring to and likewise for *"this"* in the sentence *"In most homes, this is heating/cooling and your refrigerator."* Furthermore, the following claim has one verifiable proposition about electric blankets but the remainder of the claim is difficult to verify: *"It got so cold that water turned to ice lol, best suggestion from experience is blankets plus halogen heaters that's electric as they're cheap to run, cost me about £2 a week."*

- **conditionally true:** the claim makes assumptions about cost, time, etc. or an unlikely state like needing to buy an expensive new appliance or having a specific apartment configuration. For example, the claim *"If you need to buy a new electrical appliance, for example a fridge or a washing machine, make sure that it is graded A."* is only applicable if the recipient was already planning on buying a large appliance. Additionally, the following claim is also not applicable as most apartments do not have electric water heaters: *"My home is all electric, so when I needed a new hot water heater a few years ago, I chose an electric heat pump water heater."* Lastly, the claim may be overly specific and claim a specific amount of money can be saved: *"Worse, by the time you get two chest freezers, one for fridging and one for freezing...at roughly $28 worth of electricity per year for each, your net energy savings is* *zero."*

As our experiments required using these claims as part of a feedback system, we allowed the annotators to re-write claims for style, to correct typos, or to resolve appropriate references.[23] For example, one annotator re-wrote the first-person experiential claim *"As well, I turned up the temperature in my fridge to 5C, the minimum temperature to keep food fresh, reducing how often my fridge will have to run."* by converting the unit of temperature and re-writing as the directive *"Turn up the temperature in your fridge to 41F, the minimum temperature to keep food fresh, reducing how often the fridge will have to run."* We included all tips annotated as true or as conditionally true if the claim was re-written to include the necessary context. In total, the experts annotated 63 claims, of which 48 were true after being re-written. Of the remaining claims, only three were false, showing that in this domain, false information is rare. While this dataset is too small for automated fact-checking, this number is sufficient for our experiments as we are only interested in determining the effect of the content on recipients. Future work could increase the number of claims and conduct additional experiments where we classify the claims by style, sentiment, framing, and other aspects and then evaluate their impact in a similar controlled experiment.

In addition to the expert-annotated tips, we conduct a second round of annotations in order to determine whether the annotation process could be scaled by using non-experts and in order to obtain additional data for machine learning experiments. We recruited three volunteers with backgrounds in natural language processing and computer science, all graduate students at Columbia University. The annotators were provided with the same instructions and claims as the non-experts, but were also provided access to Google search results. In total, 342 claims were annotated. As with the expert-annotated claims, false information was rare – only 7 claims were labeled as "false" – with 47 true, 57 partially true, and 35 conditionally true. Because we allowed the expert annotators to re-write conditionally and partially true claims, we were not able to compute the inter-annotator agreement on the five-way labeling task. However, the IAA across the three non-expert annotators was 0.1667, considered low agreement. The low score was mostly due to the confusion

[23] Future work could replace this process with a controlled generation approach to automatically generate stylistic variations of paraphrases.

between the three types of true labels. When we combine the three true labels into one category and "false" and "irrelevant" into another category,[24] the performance on this binary task increases to 0.7213. Framing the task as a binary task also obtains 0.6721 IAA when including the expert annotators, suggesting that modeling the task as a five-way task may be difficult and better guidelines are needed at the minimum.

6.3.2 Experiments and Methodology

We conduct two experiments on our labeled data from Section 6.3.1. For the first experiment, we evaluate several state-of-the-art fact-checking models on the 342 non-expert tips. For the second experiment, we evaluate the impact of the 48 expert-verified tips on electricity consumption behavior.

Performance of Veracity Models To determine the current capability to accurately predict the veracity of claims containing causal relations in this domain, we evaluate a number of different methods. We first consider **Our System**, which was trained on the FEVER 1.0 claims and Wikipedia evidence described in Section 6.2.1 and is thus out-of-domain on this task. Due to the domain specificity of this task, we observed that the retrieved evidence from Wikipedia is often irrelevant to the task. We thus evaluate another method, **Our System-Web Documents**, which uses all Google search results instead of only Wikipedia articles, but is otherwise the same. Consequently, this set of documents is the same set presented to the non-experts during the annotation process. Finally, we consider a model that performed well on two similar datasets for rumor detection and fact-checking in a community question answering forum [Karadzhov et al., 2017]. Both of these datasets have only a few hundred examples, similar to our task. The system of Karadzhov et al. [2017] uses an **LSTM** with input from the claim, Google search results, and Bing search results.

We conduct these experiments as a binary task, predicting "true" or "not true" as discussed in Section 6.3.1, setting aside 50 examples (25 of each) for a balanced evaluation set. We tune

[24] Because there are not enough "false" claims for modeling

hyper-parameters for the **LSTM** model by setting aside 50 additional examples and using the same settings as the authors. We find that both **Our System** and **Our System-Web Documents**, which are not re-trained on the new labeled data, are not able to generalize to this domain and always predict "not enough information," which corresponds to "not true." In a way, the performance is not surprising, as claims in the FEVER dataset are often much simpler than the energy-saving tips in terms of sentence structure, sentence length and grammar. Thus, our model is unable to generalize without fine-tuning and re-training. We also found that the **LSTM** model performs better, but still performs relatively poorly, with only 56% accuracy on the evaluation. However, with additional training data, we might expect these models to perform better. As this task remains challenging for computational models, we use the expert-verified tips for all remaining experiments.

Impact of Tips Our behavioral experiments on the effect of the 48 manually-verified tips were carried out as part of a larger experiment on electricity consumption in Columbia housing. Several Columbia buildings were outfitted with sensors to allow for the collection of data on the electricity usage of individual apartments. Each apartment has a separate meter that records the cumulative kilowatt-hours consumed in 15 second intervals and sends this information over the internet where it is stored in a database. The consumption per apartment over various intervals can then be retrieved and sent via e-mail to participants in the experiment. Participants may receive information about the current electricity consumption, a comparison to similar apartments or a previous time period, or our tips described in Section 6.3.1; this allows us to address the goal of the larger experiment to determine the effect of feedback messages on electricity consumption. In order to send feedback messages, we recruited participants by approaching residents in the lobby of the building and asking them if they would like to receive feedback on their personal electricity consumption.[25] Tenants were also informed that they could unsubscribe from the e-mail list at any time and, in fact, several initial participants selected this option. Initially, 88 apartments signed up to receive feedback. During the course of the experiment, 17 participants elected to unsubscribe via e-mail,

[25] The participants were not informed they were part of an experiment in order to prevent the results from being biased. This study was approved by the Columbia Institutional Review Board under IRB Protocol Number AAAR1391(M00Y03).

leaving 71 apartments for analysis. The control group totaled the remaining 313 apartments.

Participants in the experimental group received feedback messages every Monday and Friday according to a template-based system. Messages varied along 10 dimensions, or features, such as the sentiment of the message or the unit of power for the electricity consumed. These features were mostly independent, although included some constraints such as that when the comparison to a previous time period is included, if the electricity consumption increased and positive sentiment was selected, sentiment would be changed to neutral. Finally, given the features and constraints, the message was realized according to a number of pre-defined phrases.

The full set of features is listed in Table 6.17. Most features are binary (enabled or disabled) but others are categorical. The graph is a visualization of the electricity consumption over the past three days, in one-hour intervals, and is always included. If the "sentiment" feature is non-neutral, the message inserts one of a small set of positive phrases (e.g. *"congratulations"* or *"good job"*) or negative phrases (e.g. *"unfortunately"* or *"worse"*). "Power unit" refers to the expression for the total electricity consumption of the current feedback period (the most recent three days) and may be in kilowatt-hours as provided by the apartment meters or converted to another unit such as money or greenhouse gas emissions. "Comparison to peers" and "comparison to self," if included, refer to the percentage of electricity consumption increase or decrease relative to similar-size apartments or the previous feedback period, respectively. "Peak time information" informs the recipient if they used more electricity during the day time (between 9 am and 6 pm) or night time (between 6 pm and 9 am). The purpose of this feature is to encourage the recipient to transition high-electricity activities (such as running the dishwasher) to the night time, when the electricity grid has fewer people using it and is therefore less "stressed." The "energy-saving tip" feature indicates whether one of the tips described in Section 6.3.1 is included. Finally, "phantom load" refers to whether we include the total phantom load, which is reported in watts, the equivalent number of light bulbs, and the money that could be saved in the course of a year if this consumption was halved. The phantom load is computed as the 1st percentile[26] electricity consumption at the 15 second interval,

[26] Using the minimum is another possibility, although the 1st percentile allows for errors in the metering process that result in zero consumption being recorded

Feature	Options
Graph	{True}
Sentiment	{Positive, Neutral, Negative}
Power Unit	{kilowatt-hours, trees cut down, CO2 emissions, greenhouse gas emissions, car miles driven}
Comparison to Peers	{True, False}
Comparison to Self	{True, False}
Peak Time Information	{True, False}
Energy-Saving Tip	{True, False}
Phantom Load Information	{True, False}

Table 6.17: Features used in feedback messages and their possible values

in order to approximate the appliances that are "always on."

Each feature was randomly selected, except for whether the participant received an energy-saving tip. Instead, we divided the experiment group into three sub-groups: those that never received a tip, those that always received a tip, and those that randomly received a tip. Given that the "tip feature" was selected, we randomly selected a tip from the collected and annotated data described in Section 6.3.1, subject to two additional constraints. The first constraint is seasonal, where we restrict the pool of tips according to the likely temperature. In other words, if the current date is between 10/1 and 4/1 then we don't send any hot-weather tips such as those related to fans, windows, or air conditioning. Likewise, if the current date is between 5/1 and 11/1 then we don't send any cold-weather tips such as those related to space heaters or electric blankets. The second constraint is due to the "phantom load feature." If this feature is enabled, then we always send a tip related to the phantom load, regardless of whether the tip feature is enabled or not,[27] in order to make the message more coherent. We also include the constraint that a tip is only received once, which overrides any other constraints. Examples of feedback messages are presented in Figures 6.9 and 6.9.

[27] Only for the two sub-experiment groups that may receive tips

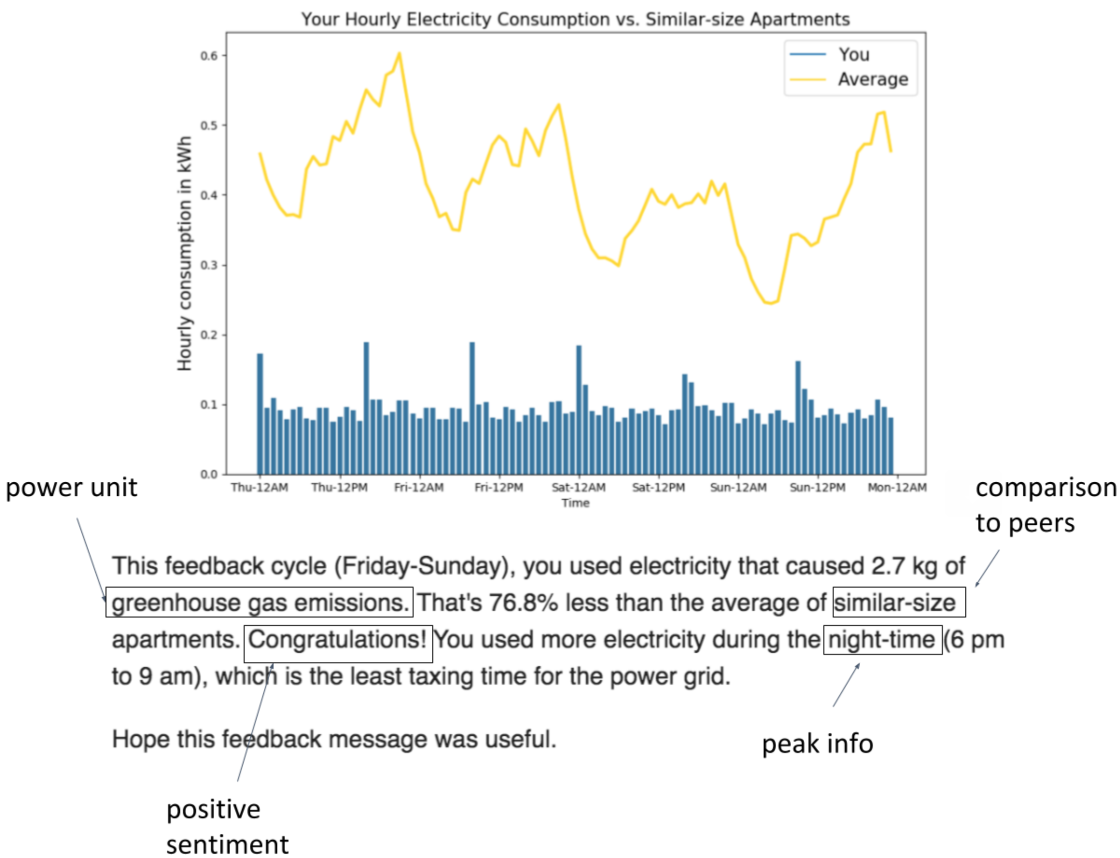


Figure 6.9: An example of a feedback e-mail that does not contain an energy-saving tip, diagrammed with the features and their lexicalization.

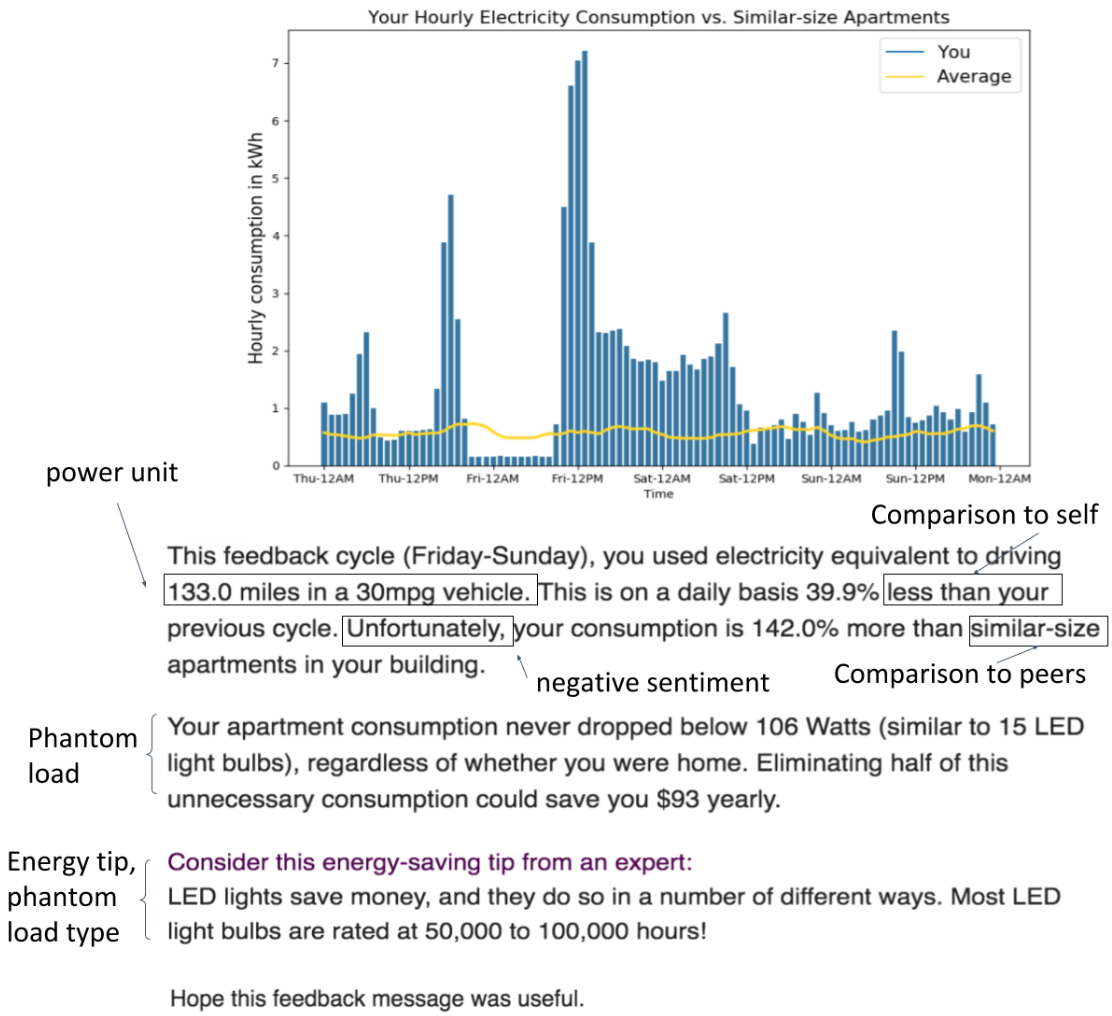


Figure 6.10: An example of a feedback e-mail that contains an energy-saving tip, diagrammed with the features and their lexicalization.

6.3.3 Results

Our resulting dataset consists of feedback messages and the corresponding power consumption between the dates of 2018/10/23 and 2019/01/31. We assume that power consumption is distributed log-normally and discarded those apartments whose mean consumption during this time period fell outside of 2 standard deviations of the mean consumption for the entire dataset. This resulted in 2 outlier apartments removed from the experiment group and 10 outliers removed from the control group, leaving 71 and 303, respectively, to be used in all subsequent analysis.

Measures of Consumption Reduction As we are interested in the effect of our energy-saving suggestions, we need a way to compare the reduction in electricity across our 3 control groups. We use the "baseline-adjusted comparison" method of Meinreken [2020], where the total reduction is compared to a period of time (the baseline) before the experiment began. Part of this equation is computed as the ratio of the mean electricity consumption of the experiment group during the baseline period divided by the mean of the control group during the same period. A similar ratio is computed for the feedback period (the entire timeframe during which participants received messages). Then, if the second ratio is smaller, we can observe that the experiment group reduced their electricity relative to the baseline (adjusted for their initial position relative to the control group). The baseline adjustment is computed according to the following equation:

$$\frac{\frac{1}{|A_e|}\sum_{a_e\in A_e}\frac{1}{|T_b|}\sum_{t_b\in T_b}P_{a_e,t_b}}{\frac{1}{|A_c|}\sum_{a_c\in A_c}\frac{1}{|T_b|}\sum_{t_b\in T_b}P_{a_c,t_b}}-\frac{\frac{1}{|A_e|}\sum_{a_e\in A_e}\frac{1}{|T_f|}\sum_{t_f\in T_f}P_{a_e,t_f}}{\frac{1}{|A_c|}\sum_{a_c\in A_c}\frac{1}{|T_f|}\sum_{t_f\in T_f}P_{a_c,t_f}} \tag{6.16}$$

where A_e and A_c are the set of apartments in the experiment group and control group, respectively, and T_b and T_f are the set of dates in the baseline and feedback periods, respectively. $P_{a,t}$ is then the total power for a given apartment on a specific date. For a baseline, we used the dates between 2018/10/01 and 2018/10/22, before the experiment started and after the time period where participants would use air conditioning, as including those days may cause high variance in the dataset. The mean daily kilowatt-hours (KwH) of the control group was 7.73 KwH during the baseline and

7.7 KwH during the feedback period, suggesting that there were few exogenous factors affecting electricity consumption during this time. However, the experiment group reduced their consumption from 9.21 KwH during the baseline to 8.33 KwH during the feedback, for a baseline-adjusted reduction of 10.89%.

Measures of Tip Impact While this analysis shows that the feedback messages are successful, the hypothesis of interest is whether the energy-saving tips lead to positive behavior change. The reduction in electricity may be due to specific features in the feedback message or simply by signing up for the experiment and increasing awareness of one's own consumption. We thus desire a measure of identifying the *additional* improvement over receiving feedback messages. One way to compute this measure is by adjusting the baseline period to instead include the first several days of the experiment and comparing differences between the three groups of tip recipients. During this period, the participants have received messages but have received very few tips. Thus, a reasonable assumption is that while they may reduce their consumption relative to the baseline during this period, if they reduce their consumption further, it may be because some participants received advice on how specifically to reduce their consumption, resulting in persistent behavior change, whereas others did not.

We set the new baseline to be the dates between 2018/10/23 and 2018/10/31, inclusive. We then compute three experiment end dates. First, we consider the date 2018/12/14, the last day of the Fall 2018 semester, as some apartments may become unoccupied for the holidays and skew the results. However, for comparison we also use the end of experiment date 2019/01/31, as while some apartments may be unoccupied, the participants that received tips may consequently take actions such as unplugging appliances. Finally, we report the results for the end date of 2019/04/01, when a second, unrelated experiment began. The reason for this date is to observe any persistent behavior change in the tip-receiving groups; the non-tip group may revert to old behavioral patterns in comparison. Using these dates, we compute the baseline-adjusted reduction using Equation 6.16 with A_e as each of the three groups to find any indication of behavior change.

Indeed, we find some evidence that relative to the initial feedback period, the groups that received tips maintained the initial reduction in energy consumption whereas the non-tip group did not. For the end-of-semester date, the non-tip group actually *increased* their electricity consumption relative to the initial period by 6.3%, whereas the groups that always received a tip further reduced their consumption by 4.4% while the group that sometimes received a tip maintained the same behavior (a reduction of 0.6%). We compute the statistical significance of this result using the paired t-test for unequal variance between the tip and non-tip group and we find significance at the $p = 0.04098$ level. For the end-of-experiment time period, we find that the non-tip group again increased their consumption, this time by 4.7%, with the always-tip and sometimes-tip groups reducing consumption by 3.4% and 3.3%, although this result is not significant ($p = 0.148$). As this period includes the semester break, the result may be skewed by empty apartments. Similarly, for the last time period, we find that the non-tip group increased their consumption by 6.9%, with the always-tip group reducing by 3.5% and the sometimes-tip group by 4.5% (not significant at $p = 0.088$.

To visualize these results, in Figure 6.11 we present the reduction computed over every date between 2018/11/12 and 2019/04/01. Note that at the end of 2018, all groups reduce their electricity consumption, reducing the impact of the tips. Also note that the non-tip group gradually increases their consumption as the time since the end of the experiment increases. In contrast, both groups that received tips approach the same overall reduction at the end of experiment and maintain that reduction after the experiment, suggesting that the knowledge resulting from the tips was a persistent effect.

6.3.4 Conclusions, Limitations, and Future Work

We presented a method for mining suggestions for saving electricity from social media using causal relations. We used experts to label a subset of these examples for veracity and conducted experiments comparing the performance of the experts to that of non-experts with access to Google search. We also evaluated the performance of several systems on these labeled claims. Finally,

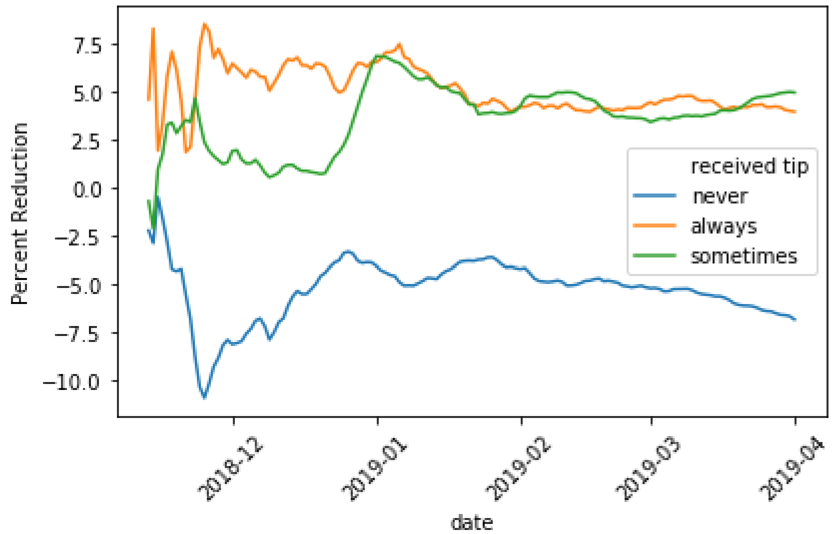


Figure 6.11: The percent reduction in electricity for the three tip sub-groups (those that never, always, or sometimes received a tip. The reduction is adjusted using Equation 6.16 and the initial experiment period between 10/23/18 and 10/31/18 inclusive.

we conducted an experiment on household electricity consumption by presenting these tips to recipients of feedback messages, finding that those that received tips reduced their consumption relative to those that did not.

While the energy-saving suggestions mined using our approach showed promise, there are a number of limitations. First, the annotation process using experts is time-consuming and non-experts, while capable, do not perform as well as experts. It is worth exploring, then, whether a distant-labeling approach is possible, such as the ones used in Sections 4.2.1 and 6.1.2. Our experts found that 25% of the tips were irrelevant and only 5% were false, suggesting that the larger corpus of unlabeled tips could be leveraged. This approach would likely require query reformulation, as many of the expert-annotated tips were re-written to be fully verified. Another avenue for obtaining data, rather than social media, would be to mine causal relations from scientific papers in electricity consumption. Claims from this domain are less likely to be experiential and as we discussed in Section 6.2.5, non-experiential claims may be easier to verify. Given the difficulty of the annotation task, future work could also explore the converse experiment: if a tip has no effect, is it likely to be false? This approach could consider the effect of individual tips rather than individual participants conditioned on receiving curated tips. The experiment would need to account for noise in the data resulting from an increase or decrease due to confounding variables (linguistic or otherwise) and would require a relatively balanced set of true and false tips.

Second, our labeled dataset does not include evidence, only claims. In order to build a system such as the one we developed in Section 6.2.3, we would require a larger dataset with evidence paired with each claim. As the annotators already used Google search to verify the evidence, this would be a small modification to the annotation task.

Finally, our results suggest that fact-checking approaches may need to handle causal relations explicitly. As constructed currently, all systems we evaluated predict the veracity of causal relations as part of an end-to-end system. However, we may need separate components to identify the direction of the causal relation (reason or result) and then evaluate each event in the causal relation individually. We may also need to evaluate these statements for beliefs or consider other forms of query normalization.

Chapter 7: Conclusions

In this thesis, we presented approaches to analyze arguments, generate counter-arguments, and make progress on some of the semantic challenges of argumentation. Our *analysis of arguments* showed that persuasion is determined in part by structure, semantics, speech acts such as agreement, and content coherence. We built models for the prediction of argumentative components, argumentative relations, and persuasive influence. For our approach to *counter-argument generation*, we applied these methods, in conjunction with models for stance detection and argument shell generation, in a hybrid retrieval-based and generation-based counter-argument system. We also presented an alternative approach, generating contrastive claims by making semantic edits. Then, we illuminated some of the *semantic challenges for argumentation* and provided solutions towards addressing these challenges: identifying causal relations by leveraging lexical indicators and improving fact-checking by creating adversarial examples. We finally applied these methods towards argument generation and analysis via a claim retrieval system in the energy domain.

Overall, this thesis contains six major contributions. The first contribution is a new dataset and methods for argumentation mining. In Section 4.1, we provided a dataset from a persuasive discussion forum (Change My View) labeled for claims and premises, the relations between them, and their semantic types. We presented an analysis of winning arguments and found three main semantic aspects: winning arguments tend to 1) begin by agreeing with their opponent 2) use a combination of pathos and logos and 3) contain coherent sequences of similar semantic types. Given this dataset, we also described in Section 4.2 how we develop a novel approach for the detection of argumentative components – claims and premises – and the relations between them. To build these models, we created two distant-labeled corpora for transfer learning – the IMHO corpus, to capture discourse context between claims and premises within a social media post, and the QR corpus, to capture dialogue context across social media posts. Then we fine-tuned large pre-trained language models on these corpora as an intermediate step, before fine-tuning on our annotated Change My View data.

The second contribution, described in Section 4.3 consists of a model for persuasive influence detection. This approach models the sequencing of sentences in a social media post using a recurrent neural network over words, frames, and discourse relations along with the interaction with another post in the thread using word overlap and a memory network. We also provided analysis showing that this model performs best on more emotional topics, supporting our results from the first contribution, and that our model outperforms novice humans on this task as well as the previous state-of-the-art approach.

Third, in Section 5.1 we provided a hybrid retrieval-based and generation-based approach for providing counter-arguments. We created a new dataset for stance detection, leveraging the thread structure of Change My View discussions to obtain additional labeled data. We then provided counter-arguments using a three-stage approach: 1) predicting the stance of a retrieved argument towards a claim and supporting argument 2) extracting the most persuasive arguments among multiple posts using our methods from Section 4.2 for argument component identification and Section 4.3 for persuasive influence prediction and 3) fusing the extracted arguments using a delete and replace approach by fine-tuning a pre-trained language model. Our approach outperforms a method that obtained state-of-the-art results for machine translation and extractive summarization, due to the semantic limitations of end-to-end models.

Next, as an alternative approach to our work in Section 5.1, we developed a new distant-labeled dataset and neural methods for contrastive claim generation. We mined data from Reddit using the acronym FTFY and cleaned the data by annotating a subset and building a classifier. Then, we built on existing sequence-to-sequence with attention methods by adding embeddings for a counter and topic, an additional loss term to predict whether a token would be copied, and constrained decoding with beam search. Our approach outperforms a strong baseline – a sequence-to-sequence model with attention.

We also showed some of the semantic challenges that remain for argumentation and possible remedies to address them. In Section 6.1, we described our methods for discovering lexical indicators of causal relations and leveraging these causal indicators for prediction. We provided a new distant-labeled dataset and methods for causal relation detection, improving over baselines that perform well on a similar task. We also showed that our approach is useful for retrieving claims in the domain of household electricity consumption (Section 6.3.1).

Finally, we made progress in the area of fact-checking and discussed our work in Section 6.2. We introduced a dataset of adversarial examples for multi-hop reasoning, temporal reasoning, and lexical and entity ambiguity. Then, we addressed these issues using a pointer network, temporal post-processing, and distributed representations, respectively. Our results obtained state-of-the-art over the previous work on the FEVER dataset and we demonstrated improvement on our adversarial claims as well. We showed that our model does not generalize out of domain, however, and that additional work is needed for fact-checking in a domain-specific area such as energy usage (Section 6.3.2). Our manually fact-checked energy saving suggestions, on the other hand, resulted in a significant reduction in electricity consumption for apartments that received these tips versus those that did not (Section 6.3.3). These results suggest that automated fact-checking would help significantly for persuasive argument generation.

7.1 Limitations

Our approach has a number of limitations, some of which are intentional decisions and some of which are open questions.

First, all of our models make no assumptions about prior beliefs or personality types. While other work has shown that these aspects play a significant role in persuasion [Lukin et al., 2017, Durmus and Cardie, 2018a, 2019b, Wang et al., 2019], we assume in all our work that this information is not available for two reasons. One reason is that in our setting, users post anonymously and obtaining this information would be difficult. The second reason is due to privacy issues – in the future users may be reluctant to share this information. Instead, we use at most information posted by a user as part of a discussion. Our energy-saving suggestions were not customized to the recipient, for instance. Our work in the analysis of arguments and generation considers at most a discussion thread, although future work could consider additional context like the user's post history.

Second, our annotation and evaluation of arguments largely did not consider the role of domain expertise. We use Mechanical Turk workers for annotation of some argumentative tasks and although we used qualification tests for our work on contrastive claims, understanding of stance may improve with expert knowledge. This lack of expertise may have affected the novice crowdworkers judgments of persuasion in Section 4.3.3 and experts in the topics of discussion or in the theories of persuasion may have performed better. Additionally, while we often used Computer Science students for annotations (e.g. our annotation of argument structure in Section 4.1.1 and our evaluation of counter-argument generation in Section 5.1.4), making inferences about the quality and relations of arguments may be difficult without deep knowledge of a topic. In our annotation of energy claims for veracity, we did use experts in the domain of household electricity consumption, although this task is time-consuming for experts which is one of the advantages of using non-experts. However, future work could also analyze the agreement between novice annotators and experts in a field such as economics on arguments in that domain. While we study online debates, this aspect is especially important in other areas such as legal texts.

Third, our work assumes that the arguments we study are conducted in good faith. While this is a reasonable assumption in Change My View, questions remain about how to detect and argue against bad faith argumentation. Prior work has examined the derailment of online conversations, as discussed in Section 2.1.1. However, this work largely focused on predicting the degeneration of a conversation [Zhang et al., 2018, Chang and Danescu-Niculescu-Mizil, 2019] rather than methods for intervening or responding. Future work may be able to leverage our analysis of effective arguments even in a bad-faith scenario. Specifically, we might expect some of the same effective techniques we observed in Section 4.1.4 to also be effective in this setting. If anyone can become a "troll" due to reasons such as a negative mood [Cheng et al., 2017], providing an empathetic and *pathos*-laden response, for example, may be more insightful to one's situation than responding in kind.

Another limitation is the retrieval-based approach. While we selected this approach for a number of reasons, such as controllability, the retrieval-based approach can only provide arguments that have been discussed previously and thus have low recall. On the other hand, generation-based approaches are limited to what they have seen in the training data, even for large pre-trained models, but in theory are capable of generating new sequences of reasoning if we can improve their understanding of semantics. Work in controllable, hierarchical generation may improve the generation of arguments if we can incorporate discourse semantics, dialogue acts, and world knowledge. Additionally, it is important for models to understand similarity between arguments at a higher level. When given a topic or claim, even one that has just recently appeared in the news, it may be possible to generalize to other situations or events [Bilu et al., 2019].

7.2 Challenges and Future Work

While the focus of this work is on form and content in argumentation, experimenting with discourse connectives and argumentative shells for the former and semantic types and stance detection for the latter, for example, many challenges remain in these areas. With the rise of pre-trained language models [Peters et al., 2018, Devlin et al., 2019], recent work [Bender and Koller, 2020] has questioned the notion that these models are learning semantics given only language as training data without any form of grounding. As we make extensive use of these models [Devlin et al., 2019, Lewis et al., 2019], one question is whether any deep argument understanding is performed and another question is whether it is even necessary if argumentation is intrinsically superficial. The extent to which content matters is an open question, and whether formal domains such as legal texts or scientific publications value meaning over form is unknown. Even in the area of professional debates, Wang et al. [2017] found that style and content affected the outcome jointly. In our work, in Section 4.2.3, we found improved results by fine-tuning language models on the appropriate form (either discourse using the IMHO+content dataset or dialogue using the QR dataset). Moreover, our results in Section 4.3.4 showed that our models outperform novice humans at the task of identifying when a counter-argument is persuasive to an original poster; this is likely only possible because our models are better at identifying shallow patterns of commonly-used argument forms. We also found in Section 5.1.2 that we can retrieve content for use as a counter-argument and then make edits for form by inserting new argument shells, neither of which require a deep understanding of argumentation. Our retrieval-based approach only requires finding similar arguments used in prior discussions and ranking them according to shallow features (e.g. structure and length). For common topics, we can achieve this due to the abundance of data.

Still, while our approach to argument generation may cover a majority of topics and argumentative strategies, the gap between current performance and human-level performance is significant. Our work often leverages lexical cues (e.g. in Section 4.2.2) for identifying argumentative components and relations and thus our modeling is non-contextual and does not require a deep understanding of argumentation. However, argumentation is inherently *contextual*. Consider the statement *"7 people died during Ebola vaccine tests."* In a news article, this would likely be a neutral, factual statement. However, if this statement was uttered in response to the claim *"Vaccine against Ebola is necessary,"* it would likely be argumentative. Saint-Dizier [2018] provides this example in support of a knowledge-based approach to understanding implicit warrants. The argument reasoning comprehension task [Habernal et al., 2018a] and follow-on work [Niven and Kao, 2019] found that state-of-the-art models such as BERT perform no better than random at this task. Current models of argumentation lack the ability to understand conversational implicature (i.e. understanding what is meant versus what was uttered [Grice, 1968]). Macagno and Walton [2013] claim that implicatures are often used as a communicative strategy in argumentation, whereby a proponent puts the burden of proof on the opponent to provide the missing link. "Common-sense reasoning," then, is necessary for better models of argumentation. Recent research in this area has resulted in large datasets and models for predicting common-sense relations [Bosselut et al., 2019] and performing inference with multiple latent steps [Bosselut and Choi, 2019] and provides a promising direction for applications to argumentation.

Additionally, future work could explore the role of established persuasive techniques from social science, which require a better understanding of meaning. The Socratic method of asking questions in a dialogue [Birkerts, 1997] has been shown to be an effective rhetorical strategy in legal reasoning [Yong, 2010] and may be useful for online argumentation as well. These questions take the form of open-ended and follow up questions [Paul and Elder, 2007]. Building on previous work in *why*-question answering [Sharp et al., 2016] may also help with *why*-question generation. Furthermore, reflecting back the other person's responses by summarizing their answers, showing empathy, and finding areas of agreement is key to reconciling opposing views in argumentation [Kroll, 2008]. Finally, providing personal narratives in support of an argument is another important technique [Hornikx, 2005, Song et al., 2016]. These aspects are challenges for automated argumentation going forward.

Other semantic relations in addition to contrast and causality are likely to be helpful for argumentation as well. While entailment has been studied in prior work in argumentation [Cabrio and Villata, 2012], revisiting the application of techniques from natural language inference may be beneficial, as recent research has yielded large corpora for many different domains and powerful pre-trained models. Improvement in discourse relation detection specifically is also likely to result in better models of argumentation. Understanding the expansion relation would be helpful for generating examples in support of a point. Temporal relations may not be helpful in many argumentative contexts, but in a legal setting they would likely be useful, as establishing a timeline of events could establish culpability. Furthermore, considering the role of discourse in argument planning is a potential productive direction. Considering a template of argument sequencing based on discourse relations and filling in content is one possible approach.

Even if persuasive argumentation can be accomplished with a primary focus on form, there are many possible improvements in that area. Our analysis (Section 4.1.4) showed that agreement and emotion were important for convincing arguments, but we did not explicitly control for these aspects during retrieval or generation. Furthermore, retrieval-based or generation-based approaches could select persuasive arguments according to which framing is most likely to be effective. Next, the interplay features of Tan et al. [2016] showed that overlap in stop words is correlated with winning arguments. These features can be considered a form of *entrainment*, an alignment between dialogue partners. Entrainment may occur at multiple levels of linguistic representation, i.e. not just words but topics or discourse relations or syllables per second in the case of speech. Entrainment has been shown to be predictive of success in other tasks [Nenkova et al., 2008, Levitan et al., 2012] and to influence conversational behavior [Levitan, 2013]. Levitan et al. [2016] found that entrainment was key for conversational avators to establish *trust*, which is necessary for *ethos* or establishing credibility, one of the weapons of influence described by Cialdini [2005]. Language may be perceived as trustworthy even when deception is used [Levitan et al., 2018], making it important to distinguish form (trusted language) from content (veracity, which may be determined by a model such as our approach in Section 6.2.3. Future work could thus experiment with re-writing arguments for entrainment or trustworthiness.

However, form and meaning are not independent. For example, the role of belief [Diab et al., 2009] in argumentation has not been thoroughly explored. The detection of hedging (non-committed belief) is important for determining the strength of a response, which is largely functional. However, the identification of reported belief may be necessary for understanding stance. An argument may be presenting multiple other viewpoints but not necessarily adopting them and leveraging belief types may help in identifying multiple sides of an issue.

Ultimately, progress in argument analysis and generation is likely to be dependent upon better models of semantics. Understanding a sequence of reasoning is necessary for identifying argument fallacies [van Eemeren and Grootendorst, 1992], which may not be necessary for online dialogues but may be important for other domains. Furthermore, many of the aforementioned semantic relations (e.g. causal relations) align with the theoretical notions of argumentative schemes or strategies [Walton, 1995].

7.3 Themes and Final Conclusions

While our work is specific to argumentation and semantics, we take a number of general approaches that we hope will transfer to other tasks.

First, we make extensive use of distant supervision. We use Internet acronyms to obtain data for discourse context (IMHO) and contrastive claims (FTFY). We leverage metadata to identify dialogue context (the QR dataset) and naturally-labeled data for persuasion (CMV). We also take advantage of the constraints imposed by the moderators of CMV that require that every response is a counter-argument to identify stance and we leverage the constraints of Simple and English Wikipedia to identify paraphrases. However, while resources like Reddit and Wikipedia make this data extensively available, we still need methods to clean and process the data for phenomena of interest. We thus annotate data or leverage pre-trained language models to reduce noise and improve quality.

Second, we take advantage of lexical indicators for difficult semantic tasks. We create approaches to discover new causal connectives which improves precision and recall of causal relation detection. We train a model to recover argumentative shells in order to improve the fluency of transitions between arguments. We also show that fine-tuning on datasets like IMHO can improve performance on claim and premise detection even without considering the surrounding context – suggesting that the model is learning discriminative phrases of claims.

Finally, we show that non-lexical and non-neural features can be helpful and can even be combined with neural models. Our models for persuasive influence combine features for length and structure, among others, with contextual representations of posts. Our models for causal relations use features based on the entire corpus – the KL-divergence of conditional distributions. Even state-of-the-art methods such as BERT can be aided by features based on RST structure and temporal relations (e.g. in Sections 4.2.2 and 6.2.3).

We view our work in this thesis as a step towards further understanding of arguments. While we focused specifically on online debates we hope that the data and methods developed are helpful for other domains and tasks. We also hope that our work on the retrieve-and-edit approach to generation leads to improvement in end-to-end models, towards the end goal of fully controllable generation of arguments. We have made available much of the code and data created during the course of this thesis: `http://www.cs.columbia.edu/~chidey/`. For any additional questions, please contact the author.

Bibliography

Mark Aakhus, Smaranda Muresan, and Nina Wacholder. An argument-ontology for a response-centered approach to argumentation mining. 2014.

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 2–11, Portland, Oregon, June 2011. Association for Computational Linguistics.

Pablo Accuosto and Horacio Saggion. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy, August 2019. Association for Computational Linguistics.

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. SemEval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California, June 2016. Association for Computational Linguistics.

Ameeta Agrawal, Aijun An, Papagelis Chen, and Manos. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2018.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling frames in argumentation. 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 2922–2932, Hong Kong, China, November 2019. Association for Computational Linguistics.

Khalid Al Khatib, Hinrich Schütze, and Cathleen Kantner. Automatic detection of point of view differences in Wikipedia. In *Proceedings of COLING 2012*, pages 33–50, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California, June 2016a. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan, December 2016b. The COLING 2016 Organizing Committee.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. Modeling deliberative argumentation strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Mohammed Alliheedi, Robert E. Mercer, and Robin Cohen. Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123, Florence, Italy, August 2019. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 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, Portland, Oregon, June 2011. Association for Computational Linguistics.

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. Giving BERT a calculator: Finding operations and arguments with reading comprehension. 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 5947–5952, Hong Kong, China, November 2019. Association for Computational Linguistics.

Jean-Claude Anscombre and Oswald Ducrot. *L'argumentation dans la langue*. Editions Mardaga, 1983.

Farida Aouladomar and Patrick Saint-Dizier. Towards generating procedural texts: An exploration of their rhetorical and argumentative structure. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, 2005.

Denis Apothéloz, Pierre-Yves Brandt, and Gustavo Quiroz. The function of negation in argumentation. *Journal of Pragmatics*, 19:23–38, 01 1993.

Aristotle, W.R Roberts, Ingram Bywater, and Friedrich Solmsen. *Rhetoric*. 1954.

David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. What gets echoed? understanding the "pointers" in explanations of persuasive arguments. 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 2911–2921, Hong Kong, China, November 2019. Association for Computational Linguistics.

Katie Atkinson and Trevor JM Bench-Capon. Argument schemes for reasoning about the actions of others. In *COMMA*, pages 71–82, 2016.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4684–4696, Hong Kong, China, November 2019. Association for Computational Linguistics.

Tal August, Nigini Oliveira, Chenhao Tan, Noah A. Smith, and Katharina Reinecke. Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. *PACMHCI*, 2(CSCW):22:1–22:19, 2018.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 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, April 2017a. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics.

Roy Bar-Haim, Dalia Krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilu, Alon Halfon, Yoav Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 977–990, Florence, Italy, July 2019. Association for Computational Linguistics.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75, Berlin, Germany, August 2016. Association for Computational Linguistics.

Beata Beigman Klebanov, Binod Gyawali, and Yi Song. Detecting good arguments in a non-topic-specific way: An oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 244–249, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Emily Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. preprint under review, 2020.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. A taxonomy of argumentation models used for knowledge representation. *Artif. Intell. Rev.*, 33(3):211–259, 2010.

Yonatan Bilu and Noam Slonim. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany, August 2016. Association for Computational Linguistics.

Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93, Denver, CO, June 2015. Association for Computational Linguistics.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026, Florence, Italy, July 2019. Association for Computational Linguistics.

Or Biran and Kathleen McKeown. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Or Biran and Kathleen McKeown. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168, Washington, DC, USA, 2011a. IEEE Computer Society. ISBN 978-0-7695-4492-2.

Or Biran and Owen Rambow. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing*, 5(4):363–381, 2011b.

Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45, Montréal, Canada, June 2012. Association for Computational Linguistics.

Sven Birkerts. The socratic method. 1997.

Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. Building and refining rhetorical-semantic relation models. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 428–435, Rochester, New York, April 2007. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. Toward stance classification based on claim microstructures. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 74–80, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Antoine Bosselut and Yejin Choi. Dynamic knowledge graph construction for zero-shot commonsense question answering. *ArXiv*, abs/1911.03876, 2019.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Peter Bourgonje and Olha Zolotarenko. Toward cross-theory discourse relation annotation. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 7–11, Minneapolis, MN, June 2019. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Chloé Braud and Pascal Denis. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213, Austin, Texas, November 2016. Association for Computational Linguistics.

Susan E. Brennan. Lexical entrainment in spontaneous dialog. 1996.

Dominique Brossard. New media landscapes and the science information consumer. *Proceedings of the National Academy of Sciences of the United States of America*, 110 Suppl 3:14096–101, 2013.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

Kasia Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yakorska. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 917–924, Reykjavik, Iceland, May 2014a. European Language Resources Association (ELRA).

Katarzyna Budzynska and Chris Reed. Speech acts of argumentation: Inference anchors and peripheral cues in dialogue. In *Computational Models of Natural Argument*, 2011.

Katarzyna Budzynska and Chris Reed. The structure of ad hominem dialogues. In *COMMA*, 2012.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. Towards argument mining from dialogue. In *COMMA*, 2014b.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. Theoretical foundations for illocutionary structure parsing. *Argument and Computation*, 7:91–108, 2016.

Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Elena Cabrio, Sara Tonelli, and Serena Villata. From discourse analysis to argumentation schemes and back: Relations and differences. In João Leite, Tran Cao Son, Paolo Torroni, Leon van der Torre, and Stefan Woltran, editors, *Computational Logic in Multi-Agent Systems*, pages 1–17, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40624-9.

Charles B. Callaway and James C. Lester. Narrative prose generation. In *Artif. Intell.*, 2001.

Amparo Elizabeth Cano-Basave and Yulan He. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California, June 2016. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver, August 2017. Association for Computational Linguistics.

Lucas Carstens and Francesca Toni. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, June 2015. Association for Computational Linguistics.

Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 127–131, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79, Florence, Italy, August 2019a. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China, November 2019c. Association for Computational Linguistics.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. 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 4742–4753, Hong Kong, China, November 2019. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017a. Association for Computational Linguistics.

Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu. Hybrid neural attention for agreement/disagreement inference in online debates. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 665–670, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017b. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language*

Generation, pages 79–88, Tilburg University, The Netherlands, November 2018b. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Zhigang Chen, Wei Lin, Qian Chen, Xiaoping Chen, Si Wei, Hui Jiang, and Xiaodan Zhu. Revisiting word embedding for contrasting meaning. 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)*, pages 106–115, Beijing, China, July 2015. Association for Computational Linguistics.

Justin Cheng, Michael S. Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. *CSCW : proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, 2017:1217–1230, 2017.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. Hedge detection as a lens on framing in the GMO debates: A position paper. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 70–79, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics.

HongSeok Choi and Hyunju Lee. GIST at SemEval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Dennis Lee Chong and James N. Druckman. Framing theory. 2007.

Robert B Cialdini. Influence: The psychology of persuasion collins. *Revised edition (October 7, 2005)*, 2005.

Oana Cocarascu and Francesca Toni. A system for supporting the detection of deceptive reviews using argument mining. In *COMMA*, 2016.

Oana Cocarascu and Francesca Toni. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

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

Alexander Coppock. Positive, small, homogeneous, and durable: Political persuasion in response to information. 2016.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *MLCW*, 2005.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon M. Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *WWW '09*, 2009.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan T. Dumais. Mark my words!: linguistic style accommodation in social media. In *WWW*, 2011a.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon M. Kleinberg. Echoes of power: language effects and power differences in social interaction. In *WWW*, 2011b.

Abhimanyu Das, Sreenivas Gollapudi, Emre Kıcıman, and Onur Varol. Information dissemination in heterogeneous-intent networks. In *Proceedings of the 8th ACM Conference on Web Science*, pages 259–268. ACM, 2016.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, California, June 2010. Association for Computational Linguistics.

Clarissa C. David, Ma. Rosel S. San Pascual, and Ma. Eliza S. Torres. Reliance on facebook for news and its influence on political engagement. In *PloS one*, 2019.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcasm in twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Paul Deane. On the relation between automated essay scoring and modern views of the writing construct. 2013.

Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Meeting of the North American Association for Computational Linguistics*, 2019.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Tao Ding and Shimei Pan. Personalized emphasis framing for persuasive message generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1432–1441, Austin, Texas, November 2016. Association for Computational Linguistics.

Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133, 2017.

Esin Durmus and Claire Cardie. 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, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. Understanding the effect of gender and stance in opinion expression in debates on "abortion". In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 69–75, New Orleans, Louisiana, USA, June 2018b. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. Modeling the factors of user success in online debate. In *The World Wide Web Conference*, WWW '19, pages 2701–2707, New York, NY, USA, 2019a. ACM. ISBN 978-1-4503-6674-8.

Esin Durmus and Claire Cardie. 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*, pages 602–607, Florence, Italy, July 2019b. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy, July 2019a. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. The role of pragmatic and discourse context in determining argument impact. 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 5667–5677, Hong Kong, China, November 2019b. Association for Computational Linguistics.

Rory Duthie and Katarzyna Budzynska. A deep modular rnn approach for ethos mining. In *IJCAI*, 2018a.

Rory Duthie and Katarzyna Budzynska. Classifying types of ethos support and attack. In *COMMA*, 2018b.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining ethos in political debate. In *COMMA*, 2016.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy, July 2019. Association for Computational Linguistics.

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

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Corpus wide argument mining – a working solution, 2019.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium, October 2018. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan, October–November 2019. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *ACL*, pages 987–996, 2011.

Kate Forbes-Riley, Fan Zhang, and Diane Litman. Extracting PDTB discourse relations from student essays. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 117–127, Los Angeles, September 2016. Association for Computational Linguistics.

Diane Francis and Full Fact. Fast and furious fact check challenge., 2016.

James B. Freeman. Dialectics and the macrostructure of arguments: A theory of argument structure. 1991.

James B Freeman. What types of statements are there? *Argumentation*, 14(2):135–157, 2000.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1274–1290, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

Raul Garreta and Guillermo Moncecchi. Learning scikit-learn: Machine learning in python. 2013.

Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170, 2017.

Debela Gemechu and Chris Reed. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526, Florence, Italy, July 2019. Association for Computational Linguistics.

Jürgen Gerhards and Mike S. Schäfer. Is the internet a better public sphere? comparing old and new media in the usa and germany. *New Media & Society*, 12:143–160, 2010.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany, August 2016. Association for Computational Linguistics.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, SaarbrÃ¼cken, Germany, August 2017. Association for Computational Linguistics.

Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

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

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, 2018.

Trudy Govier. *A Practical Study of Argument*. Cengage Learning, Wadsworth, 2010.

Alex Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

Heather Graves, Roger Graves, Robert Mercer, and Mahzereen Akter. Titles that announce argumentative claims in biomedical research articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Lucas Graves. Understanding the promise and limits of automated fact-checking. Technical report, Reuters Institute, University of Oxford, 2018.

Nancy Green. Generation of biomedical arguments for lay readers. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 114–121, Sydney, Australia, July 2006. Association for Computational Linguistics.

Nancy Green. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24:181–196, 2010.

Nancy Green. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Nancy Green. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21, Denver, CO, June 2015. Association for Computational Linguistics.

Nancy Green. Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, pages 105–110, Brussels, Belgium, November 2018. Association for Computational Linguistics.

H. Paul Grice. Utterer's meaning, sentence-meaning, and word-meaning. 1968.

Marco Guerini, Oliviero Stock, and Massimo Zancanaro. Persuasion models for intelligent interfaces. 2003.

Marco Guerini, Oliviero Stock, and Massimo Zancanaro. Persuasive strategies and rhetorical relation selection. 2004.

Marco Guerini, Oliviero Stock, and Massimo Zancanaro. A taxonomy of strategies for multimodal persuasive message generation. *Applied Artificial Intelligence*, 21:99–136, 2007.

Chinnappa Guggilla, Tristan Miller, and Iryna Gurevych. CNN- and LSTM-based claim classification in online user comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2740–2751, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Ivan Habernal and Iryna Gurevych. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, August 2016a. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 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, November 2016b. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

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

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. 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 386–396, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China, November 2019. Association for Computational Linguistics.

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. Issue framing in online discussion fora. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1401–1407, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 619–630, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1803–1812, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4887-4.

He He, Nanyun Peng, and Percy Liang. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Jack Hessel and Lillian Lee. Something's brewing! early prediction of controversy-causing posts from discussion features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1648–1659, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy, August 2019. Association for Computational Linguistics.

Christopher Hidey and Mona Diab. Team SWEEPer: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 150–155, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Christopher Hidey and Kathleen McKeown. Persuasive influence detection: The role of argument sequencing. In *AAAI*, 2018.

Christopher Hidey and Kathy McKeown. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany, August 2016. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1756–1767, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Colin Higgins and Robyn Walker. Strategies of persuasion in social/environmental reporting. 2010.

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

Jos Hornikx. A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. 2005.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 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, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 328–339, 2018.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

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

Xinyu Hua and Lu Wang. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. Sentence-level content planning and style specification for neural text generation. 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 591–602, Hong Kong, China, November 2019. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy, July 2019a. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics.

William J. Hutchins and Harold L. Somers. An introduction to machine translation. 1992.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. Aligning sentences from standard Wikipedia to simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Cherilyn Ireton. Journalism, 'fake news' and disinformation: handbook for journalism education and training. 2018.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China, July 2015. Association for Computational Linguistics.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez i Villodre, and Preslav Nakov. Claimrank: Detecting check-worthy claims in arabic and english. In *NAACL-HLT*, 2018.

Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,*, pages 13–24, 2014.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2224, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Shan Jiang and Christo Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW): 82:1–82:23, November 2018.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Rosé, and Graham Neubig. Attentive interaction model: Modeling changes in view in argumentation. 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 103–116, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. A cascade model for proposition extraction in argumentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 11–24, Florence, Italy, August 2019. Association for Computational Linguistics.

William Jordan. British people trust wikipedia more than the news. `https://yougov.co.uk/topics/politics/articles-reports/2014/08/09/more-british-people-trust-wikipedia-trust-news`, 2014. 2014-08-09.

Garth S. Jowett and Victoria O'Donnell. Propaganda and persuasion. 1986.

Raquel Justo, Thomas Chase Corcoran, Stephanie M. Lukin, Marilyn A. Walker, and M. Inés Torres. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowl.-Based Syst.*, 69:124–133, 2014.

Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. Detecting and explaining causes from text for a time series event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2767, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria, September 2017. INCOMA Ltd.

Nora Kassner and Hinrich Schütze. Negated lama: Birds cannot fly, 2019.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, Texas, November 2016. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas, November 2016. Association for Computational Linguistics.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2741–2749, 2016.

Youngwoo Kim and James Allan. FEVER breaker's run of team NbAuzDrLqg. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 99–104, Hong Kong, China, November 2019. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

Paul Kingsbury and Martha Palmer. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining*, pages 156–166, Florence, Italy, August 2019. Association for Computational Linguistics.

Robert A. Kowalski and Francesca Toni. Abstract argumentation. *Artificial Intelligence and Law*, 4:275–296, 1996.

Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

Barry M. Kroll. Arguing with adversaries: Aikido, rhetoric, and the art of peace. 2008.

Canasai Kruengkrai. Learning to flip the sentiment of reviews from non-parallel corpora. 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 6310–6315, Hong Kong, China, November 2019. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics.

Majid Laali and Leila Kosseim. Inducing discourse connectives from parallel texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 610–619, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. 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 3577–3582, Hong Kong, China, November 2019. Association for Computational Linguistics.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Tamar Lavee, Lili Kotlerman, Matan Orbach, Yonatan Bilu, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 29–38, Hong Kong, November 2019a. Association for Computational Linguistics.

Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 58–66, Florence, Italy, August 2019b. Association for Computational Linguistics.

John Lawrence and Chris Reed. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO, June 2015. Association for Computational Linguistics.

John Lawrence and Chris Reed. Argument mining using argumentation scheme structures. In *COMMA*, 2016.

John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, pages 1–54, 2019.

John Lawrence, Rory Duthie, Katarzyna Budzynska, and Chris Reed. Argument analytics. In *COMMA*, 2016.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org, 2014.

Rivka Levitan. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 84–90, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Rivka Levitan, Agustín Gravano, Laura Willson, SÌ†tefan BenÌ†usÌ†, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada, June 2012. Association for Computational Linguistics.

Rivka Levitan, Stefan Benus, Ramiro H. Gálvez, Agustín Gravano, Florencia Savoretti, Marián Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. In *INTERSPEECH*, 2016.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. Acoustic-prosodic indicators of deception and trust in interview dialogues. In *INTERSPEECH*, 2018.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

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

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016a. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas, November 2016b. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. 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 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Bing Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*, 2010.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Yang Liu. Fine-tune bert for extractive summarization. In *https://arxiv.org/abs/1903.10318*, 2019.

Liane Longpre, Esin Durmus, and Claire Cardie. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy, August 2019. Association for Computational Linguistics.

Luca Lugini and Diane Litman. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain, April 2017. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550, 2019.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

Fabrizio Macagno and Douglas Walton. Implicatures as forms of argument. 2013.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montréal, Canada, June 2012. Association for Computational Linguistics.

Christopher Malon. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium, November 2018. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: toward a functional theory of text. 1988.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 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, June 2014. Association for Computational Linguistics.

Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

Katerina Eva Matsa and Elisa Shearer. News use across social media platforms 2018. `http://web.archive.org/web/20190123174729/http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/`, 2018. 2018-09-10.

Kathleen R. McKeown. The text system for natural language generation: An overview. In *20th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Toronto, Ontario, Canada, June 1982. Association for Computational Linguistics.

Christoph Meinreken. Energy saving behavior in response to multi-featured feedback: Anticonformity or conformity? Technical report, 2020.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Villata Serena. Never retreat, never retract: Argumentation analysis for political speeches. In *Association for the Advancement of Artificial Intelligence*, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. In proceedings of the international conference on learning representations. 2017.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. Fact checking in community forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5309–5316, 2018.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41, 1992.

Tristan Miller, Maria Sukhareva, and Iryna Gurevych. A streamlined method for sourcing discourse-level argumentation annotations from the crowd. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1790–1796, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. A recorded debating dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018a. European Language Resources Association (ELRA).

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Amita Misra and Marilyn Walker. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France, August 2013. Association for Computational Linguistics.

Amita Misra, Shereen Oraby, Shubhangi Tandon, Sharath T. S., Pranav Anand, and Marilyn A. Walker. Summarizing dialogic arguments from social media. *CoRR*, abs/1711.00092, 2017.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-680-6.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 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, June 2016. Association for Computational Linguistics.

Gaku Morio and Katsuhide Fujita. Predicting argumentative influence probabilities in large-scale online civic engagement. In *WWW*, 2018a.

Gaku Morio and Katsuhide Fujita. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In *Proceedings of the 5th Workshop on Argument Mining*, pages 11–21, Brussels, Belgium, November 2018b. Association for Computational Linguistics.

Gaku Morio and Katsuhide Fujita. On the role of syntactic graph convolutions for identifying and classifying argument components. In *AAAI*, 2019.

Elena Musi. How did you change my view? a corpus-based study of concessions' argumentative role. *Discourse Studies*, 20(2):270–288, 2018.

Elena Musi and Mark Aakhus. Framing fracking: Semantic frames as meta-argumentative indicators for knowledge-driven argument mining of controversies. 2019.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 82–93, Berlin, Germany, August 2016. Association for Computational Linguistics.

Elena Musi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Nona Naderi and Graeme Hirst. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Ndapandula Nakashole and Tom M. Mitchell. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23, Valencia, Spain, April 2017. Association for Computational Linguistics.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Ana Laura Nettel and Georges Roque. Persuasive argumentation versus manipulation. *Argumentation*, 26(1):55–69, 2012.

Nic Newman, David A. L. Levy, and Rasmus Kleis Nielsen. Reuters institute digital news report 2015. 2014.

Huy Nguyen and Diane Litman. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO, June 2015. Association for Computational Linguistics.

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

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459, Berlin, Germany, August 2016. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*, 2018.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

Piotr Niewinski, Maria Pszona, and Maria Janicka. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China, November 2019. Association for Computational Linguistics.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, Florence, Italy, July 2019a. Association for Computational Linguistics.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, Florence, Italy, July 2019b. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics.

Jan Nuyts. Notions of (inter) subjectivity. *English Text Construction*, 5(1):53–76, 2012.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 116–126, Denver, CO, June 2015. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. A dataset of general-purpose rebuttal. 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 5590–5600, Hong Kong, China, November 2019. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

ChaeHun Park, Wonsuk Yang, and Jong Park. Generating sentential arguments from diverse perspectives on controversial topic. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 56–65, Hong Kong, China, November 2019. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Joonsuk Park, Cheryl Blake, and Claire Cardie. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 206–210, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3522-5.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *CIKM '17*, 2017.

Michael Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, Cambridge, MA, October 2010. Association for Computational Linguistics.

Richard Paul and Linda Elder. Critical thinking: The art of socratic questioning, part iii. 2007.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *IJCINI*, 7:1–31, 2013a.

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

Andreas Peldszus and Manfred Stede. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In D. Mohammed and M. Lewinski, editors, *Argumentation and Reasoned Action - Proc. of the 1st European Conference on Argumentation, Lisbon, 2015*. College Publications, London, 2016a.

Andreas Peldszus and Manfred Stede. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112, Berlin, Germany, August 2016b. Association for Computational Linguistics.

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

Chaim Perelman and Lucie Olbrechts-Tyteca. *The new rhetoric: A treatise on argumentation*. University of Notre Dame Pess, 1973.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. 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*, pages 683–691, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. Fake news challenge. 2017.

Peter Potash and Anna Rumshisky. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2475, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural*

Language Processing, pages 1364–1373, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada, June 2012a. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Who's (really) the boss? perception of situational power in written interactions. In *Proceedings of COLING 2012*, pages 2259–2274, Mumbai, India, December 2012b. The COLING 2012 Organizing Committee.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Realization of discourse relations by other means: Alternative lexicalizations. In *Coling 2010: Posters*, pages 1023–1031, Beijing, China, August 2010. Coling 2010 Organizing Committee.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Iyad Rahwan and Kate Larson. Mechanism design for abstract argumentation. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 2*, pages 1031–1038, 2008.

Iyad Rahwan, Kate Larson, and Fernando A. Tohmé. A characterisation of strategy-proofness for grounded argumentation semantics. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pages 251–256, 2009.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems, IWSDS 2017, Farmington, PA, USA, 6-9 June 2017, revised selected papers*, pages 45–52, 2017.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 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, Berlin, Germany, August 2016. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

James Ravenscroft, Anika Oellrich, Shyamasree Saha, and Maria Liakata. Multi-label annotation in scientific articles - the multi-label cancer risk assessment corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.

Chris Reed and Douglas Walton. Argumentation schemes in argument-as-process and argument-as-product. 2003.

Chris Reed and Douglas Walton. Argument schemes in dialogue. 2007.

Lena Reed, Shereen Oraby, and Marilyn Walker. Can neural generators for dialogue learn sentence planning and discourse structuring? In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 284–295, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. A computational approach for generating toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55, Denver, CO, June 2015. Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy, July 2019. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30, Metz, France, August 2013. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, 2018.

Tjitze Rienstra, Matthias Thimm, and Nir Oren. Opponent models with uncertainty for strategic argumentation. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 332–338, 2013.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni. It's a contradiction – no, it's not: A case study using functional relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Hannah Rohde and William S. Horton. Why or what next? eye movements reveal expectations about discourse direction. 2010.

Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 991–996, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4144-8.

Sara Rosenthal and Kathleen McKeown. Detecting opinionated claims in online discussions. *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37, 2012.

Sara Rosenthal and Kathleen Mckeown. Detecting influencers in multiple online genres. *ACM Transactions on Internet Technology (TOIT)*, 17(2):12, 2017.

Sara Rosenthal and Kathy McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

Charlotte Roze, Chloé Braud, and Philippe Muller. Which aspects of discourse relations are hard to learn? primitive decomposition for discourse relation classification. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 432–441, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, and Piotr Mardziel. Build it, break it, fix it: Contesting secure development. *ArXiv*, abs/1907.01679, 2016.

Kristin K. Runge, Sara K. Yeo, Michael A. Cacciatore, Dietram A. Scheufele, Dominique Brossard, Michael Xenos, Ashley A. Anderson, Doo-Hun Choi, Jiyoun Kim, Nan Li, Xuan Liang, Maria Stubbings, and Leona Yi-Fan Su. Tweeting nano: how public discourses about nanotechnology develop in social media environments. *Journal of Nanoparticle Research*, 15: 1–11, 2013.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California, 2006. Distributed with the FrameNet data.

Attapol Rutherford and Nianwen Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Patrick Saint-Dizier. A knowledge-based approach to warrant induction. In *COMMA*, 2018.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114, Beijing, China, July 2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

William Sattelberg. The demographics of reddit: Who uses the site? `https://www.techjunkie.com/demographics-reddit/`, 2019. 2019-07-22.

Tatjana Scheffler and Manfred Stede. Mapping pdtb-style connective annotation to rst-style discourse annotation. In *KONVENS*, 2016.

E. Schiappa and J.P. Nordin. *Argumentation: Keeping Faith with Reason*. Pearson Education, 2013. ISBN 9780205943784.

Allen Schmaltz, Yoon Kim, Alexander Rush, and Stuart Shieber. Adapting sequence models for sentence correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2813, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Karin Kipper Schuler, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *AAAI/IAAI*, 2000.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China, November 2019. Association for Computational Linguistics.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 138–148, Austin, Texas, November 2016. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc., 2017.

Daiki Shirafuji, Rafal Rzepka, and Kenji Araki. Debate outcome prediction using automatic persuasiveness evaluation and counterargument relations. In *LaCATODA/BtG@IJCAI*, 2019.

Eyal Shnarch, Ran Levy, Vikas Raykar, and Noam Slonim. GRASP: Rich patterns for argumentation mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1345–1350, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

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

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. "you are grounded!": Latent name artifacts in pre-trained language models, 2020.

Edwin Simpson and Iryna Gurevych. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371, 2018.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 1–8, Taipei, Taiwan, November 2017. Association for Computational Linguistics.

Maria Skeppstedt, Manfred Stede, and Andreas Kerren. Stance-taking in topics extracted from vaccine-related tweets and discussion forum posts. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 5–8, Brussels, Belgium, October 2018. Association for Computational Linguistics.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO, June 2015. Association for Computational Linguistics.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6, 2007.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. Evaluating argumentative and narrative essays using graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3350–3359, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Wei Song, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. Anecdote recognition and recommendation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2592–2602, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Dhanya Sridhar and Lise Getoor. Estimating causal effects of tone in online debates. In *IJCAI*, 2019.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 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, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. Joint models of disagreement and stance in online debate. 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)*, pages 116–125, Beijing, China, July 2015. Association for Computational Linguistics.

Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):163:1–163:21, November 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15 (1):1929–1958, January 2014.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.

Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, August 2014a. Dublin City University and Association for Computational Linguistics.

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

Christian Stab and Iryna Gurevych. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118, Berlin, Germany, August 2016. Association for Computational Linguistics.

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

Christian Stab and Iryna Gurevych. 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, April 2017b. Association for Computational Linguistics.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *ArgNLP*, 2014.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics.

Dominik Stammbach and Guenter Neumann. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China, November 2019. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 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, June 2018. Association for Computational Linguistics.

Manfred Stede. Towards assessing depth of argumentation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3308–3317, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 3267–3277, Hong Kong, China, November 2019. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc., 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

Reid Swanson, Brian Ecker, and Marilyn Walker. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624. International World Wide Web Conferences Steering Committee, 2016.

James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543*, 2019.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic argument quality assessment - new datasets and methods. 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 5624–5634, Hong Kong, China, November 2019. Association for Computational Linguistics.

Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123, Berlin, Germany, August 2016. Association for Computational Linguistics.

Stephen Toulmin. *The Uses of Argument*. 1958.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. Fine-grained argument unit recognition and classification. 2019.

Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. Ranking kernels for structures and embeddings: A hybrid preference and classification model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 897–902, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Elise van der Pol, Sharon Gieske, and Raquel Fernández. Linguistic style accommodation in disagreements. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 120–124, Berlin, Germany, August 2016. Association for Computational Linguistics.

Frans H van Eemeren and Frans Hendrik Eemeren. *Examining argumentation in context: Fifteen studies on strategic maneuvering*, volume 1. John Benjamins Publishing, 2009.

Frans H. van Eemeren and Rob Grootendorst. Argumentation, communication, and fallacies: A pragma-dialectical perspective. 1992.

M. Agnes van Rees. Discourse analysis and argumentation theory: the case of television talk. 2007.

Vladimir Vapnik. Statistical learning theory. 1998.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc., 2015.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *International Conference on Learning Represen- tations (ICLR)*, 2016.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

Ivan Vulić. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 137–143, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. Identifying semantic divergences in parallel text without annotations. 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 1503–1515, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Giovanni Da San Martino, Dora Kiesel, and Benno Stein. The impact of modeling overall argumentation with tree kernels. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2379–2389, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 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, July 2017b. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain, April 2017c. Association for Computational Linguistics.

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

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain, April 2017e. Association for Computational Linguistics.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, Santa Fe, New Mexico, USA, August 2018a. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia, July 2018b. Association for Computational Linguistics.

Claudia Wagner, David García, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *ICWSM*, 2015.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596, Montréal, Canada, June 2012a. Association for Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey, May 2012b. European Language Resources Association (ELRA).

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics.

Douglas Walton. Argumentation schemes for presumptive reasoning. 1995.

Douglas Walton, Chris Reed, and Fabrizio Macagno. Argumentation schemes. 2008.

Lu Wang and Claire Cardie. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland, June 2014a. Association for Computational Linguistics.

Lu Wang and Claire Cardie. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, Maryland, June 2014b. Association for Computational Linguistics.

Lu Wang and Wang Ling. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California, June 2016. Association for Computational Linguistics.

Lu Wang, Hema Raghavan, Claire Cardie, and Vittorio Castelli. Query-focused opinion summarization for user-generated content. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1660–1669, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232, 2017.

William Yang Wang. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Jingfeng Yang. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, 2018.

Zhongqing Wang and Yue Zhang. Opinion recommendation using a neural model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1627–1638. Association for Computational Linguistics, 2017a.

Zhongqing Wang and Yue Zhang. Opinion recommendation using a neural model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1637, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany, August 2016. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280, June 1989.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. *ArXiv*, abs/1610.08914, 2016.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2397–2406. JMLR.org, 2016.

Wenting Xiong and Diane Litman. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Toshihiko Yanase, Toshinori Miyoshi, Kohsuke Yanai, Misa Sato, Makoto Iwayama, Yoshiki Niwa, Paul Reisert, and Kentaro Inui. Learning sentence ordering for opinion generation of debate. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 94–103, Denver, CO, June 2015. Association for Computational Linguistics.

Chao Yang, Shimei Pan, Jalal Mahmud, Huahai Yang, and Padmini Srinivasan. Using personal traits for brand preference prediction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 86–96, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489, 2016.

Taha Yasseri, András Kornai, and János Kertész. A practical approach to language complexity: A wikipedia case study. In *PloS one*, 2012.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Choi Min Yong. The socratic method and legal education. 2010.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.

Fangzhou Zhai, Vera Demberg, Pavel Shkadzko, Wei Shi, and Asad Sayeed. A hybrid model for globally coherent story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 34–45, Florence, Italy, August 2019. Association for Computational Linguistics.

Fan Zhang, Diane Litman, and Katherine Forbes Riley. Inferring discourse relations from PDTB-style discourse labels for argumentative revision classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2615–2624, Osaka, Japan, December 2016a. The COLING 2016 Organizing Committee.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California, June 2016b. Association for Computational Linguistics.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. Asking too much? the rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1572, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy, July 2019. Association for Computational Linguistics.

Ingrid Zukerman, Richard McConachy, and Sarah George. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62, Mitzpe Ramon, Israel, June 2000. Association for Computational Linguistics.