DISCOURSE-LEVEL RELATIONS FOR OPINION ANALYSIS

by

Swapna Somasundaran

MSE, Johns Hopkins University, 2002

Submitted to the Graduate Faculty of

the School of Arts and Sciences, Department of Computer Science

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Swapna Somasundaran

It was defended on

March 24th 2010

and approved by

Dr. Janyce Wiebe, Computer Science Department, University of Pittsburgh

Dr. Diane Litman, Computer Science Department, University of Pittsburgh

Dr. Rebecca Hwa, Computer Science Department, University of Pittsburgh

Dr. William Cohen, Machine Learning Department, Carnegie Mellon University

Dissertation Director: Dr. Janyce Wiebe, Computer Science Department, University of Pittsburgh

DISCOURSE-LEVEL RELATIONS FOR OPINION ANALYSIS

Swapna Somasundaran, PhD

University of Pittsburgh, 2010

Opinion analysis deals with subjective phenomena such as judgments, evaluations, feelings, emotions, beliefs and stances. The availability of public opinion over the Internet and in face to face conversations, coupled with the need to understand and mine these for end applications, has motivated a great amount of research in this field in recent times. Researchers have explored a wide array of knowledge resources for opinion analysis, from words and phrases to syntactic dependencies and semantic relations.

In this thesis, we investigate a discourse-level treatment for opinion analysis.

In order to realize the discourse-level analysis, we propose a new linguistic representational scheme designed to support interdependent interpretations of opinions in the discourse. We adapt and extend an existing subjectivity annotation scheme to capture discourse-level relations in a multi-party meeting corpus. Human inter-annotator agreement studies show that trained human annotators can recognize the elements of our linguistic scheme.

Empirically, we test the impact of our discourse-level relations on fine-grained polarity classification. In this process, we also explore two different global inference models for incorporating discourse-based information to augment word-based information. Our results show that the discourse-level relations can augment and improve upon word-based methods for effective fine-grained opinion polarity classification. Further, in this thesis, we explore linguistically motivated features and a global inference paradigm for learning the discourse-level relations form the annotated data.

We employ the ideas from our linguistic scheme for recognizing stances in dual-sided debates from the product and political domains. For product debates, we use web mining

and rules to learn and employ elements of our discourse-level relations in an unsupervised fashion. For political debates, on the other hand, we take a more exploratory, supervised approach, and encode the building blocks of our discourse-level relations as features for stance classification. Our results show that the ideas behind the discourse level relations can be learned and employed effectively to improve overall stance recognition in product debates.

Keywords: Opinion analysis, sentiment, arguing, linguistic scheme, annotation scheme, computational modeling, fine-grained polarity analysis, stance recognition.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

PhD is not something that I always knew I would do. However, after getting a taste of research in Natural Language Processing, the desire to pursue it as a career for the rest of my life was simply too overwhelming. So I can safely say that I went for a PhD only because I wanted to do NLP research. Thus, my heartfelt gratitude to Padmashree Professor P.V.S. Rao, for introducing NLP and opening a new world of possibilities to me, and for putting incredible faith in a novice that I was then.

The past 6 years have been an incredible journey, at academic, professional, social, personal, and sometimes, even spiritual levels. And I could not have possibly asked for a better advisor than Jan Wiebe. Jan has been my *guru* in the most complete sense of the word. Through the years, Jan has patiently guided me thought the maze of academia, encouraged me to think independently, rescued me from despair, and jubilated in my success. I could go on and on, but in the confines of this page I shall sum up with "I am truly grateful".

I have also been lucky to have incredible mentors, such as Diane Litman, Rebecca Hwa, William Cohen, Lise Getoor, David Yarowsky, Wendy Chapman, Daniel Mosse and teachers such as Greg Cooper and Milos Hauskrecht, who have provided inspiration at so many different levels. I aspire that some day I will have their tenacity, wit, clarity of thought, depth of knowledge, and wisdom. I am particularly thankful to my thesis committee, without whom my work would not have been possible.

Over these years, I have also had the good fortune of collaborating with some really nice people such as Josef Ruppenhoffer, Galileo Namata and Theresa Wilson. My work has greatly benefited from linguistic insights form Josef and machine learning insights from Galileo. Theresa as an officemate, collaborator and predecessor has shaped the directions in my thesis.

I am thankful for the camaraderie and humor of my officemates Cem Akkaya, Yaw Anti Gyamfi, Onur Cobanoglu, Paul Hoffmann and Jason Kessler. Cem, as you now ascend to our lab throne, remember that it comes with great responsibility – that of watering the lab plants. I would also like to thank the NLP group members of the University of Pittsburgh, Hua Ai, Behrang Mohit, Mihai Rotaru, Joel Tetrault, Art Ward and Wenting Xiong who have sat through and helped me with my presentations over these years.

Last but most importantly, Ankur has been a phenomenal husband and friend. He has been my lighthouse and my source of well being. I have been blessed with incredible parents and parents in-law, who have provided their support, advice and love every step of the way. Words fail me, so I shall borrow from Jane Austen and say "If I loved you less, I might be able to talk about it more".

1.0 INTRODUCTION

Subjectivity and Opinion analysis has become a burgeoning field in recent times. Subjectivity refers to the linguistic expression of speculations, evaluations, sentiments, beliefs, etc. (i.e., *private states*) [Wiebe et al., 1999, Wilson and Wiebe, 2003, Wiebe et al., 2004]. *Sentiment* is a specific subtype of subjectivity, namely linguistic expressions of evaluations, feelings and emotions[1]. Some researchers have also analyzed other types of subjectivity such as *arguing*. Arguing refers to arguing about something, arguing that something is true or should be done. In this thesis, we work on these two types of subjectivity and collectively refer to them as *opinions*.

Many different aspects of subjectivity and sentiment, such as what the subjectivity is about (the *target* of the opinion), whether the subjective expression is positive, negative or neutral (the *polarity* of the subjective expression), and who is the holder of the subjectivity (the *source* or agent) have been explored. Researchers have also studied subjectivity at different levels of granularity – at the document, sentence, expression, and word level. In this work, we propose to study opinion analysis at the level of the discourse. Discourse analysis operates at the level higher than that of a clause [2], but lower than that of a document. The analysis looks at not just what is contained in a text span, but also at relations between them.

Fine-grained opinion analysis, that is, analysis below the level of the document, is important for applications such as Question Answering (QA) to arrive at the correct answers.

[1] Some researchers also use "sentiment" to refer to the positive or negative orientation of words or documents. However, in this work, we will refer to this property as polarity.

[2] A clause is a syntactic construction containing a subject and predicate and forming part of a sentence or constituting a whole simple sentence

For instance in Example 1.1[3], there are two opinions, one positive and the other negative. In order for a QA system to answer questions regarding opinions towards iPhones (for example "Does the writer have a favorable opinion regarding the iPhone?"), it will need to delineate each opinion expression, determine if it is positive or negative, and determine what it is toward.

(1.1) The *iPhone* **incarnate the 21st century** whereas *Blackberry* symbolizes an **outdated technology.**

A popular approach in fine-grained sentiment analysis has been to identify the opinion of the individual words or expressions using lexical resources such as the General Inquirer [Stone, 2000], Wordnet [Miller, 1990] or Subjectivity Lexicon [Wilson et al., 2005a]. While these resources are indeed useful, there are still ambiguities that the lexicon-based approaches cannot easily resolve. In this work, we suggest a discourse-based approach to improve fine-grained polarity disambiguation.

We will motivate our discourse-based approach for fine-grained opinion analysis with a snippet (Example 1.2) from the AMI meeting corpus[4]. At this point in the meeting, the participants are discussing what material should be used for the TV remote. Here the speaker really likes the rubbery material and wants it to be accepted by the group.

(1.2) D:: ... this kind of rubbery material, *it's* a **bit more bouncy**, like you said they get chucked around a lot. A **bit more durable** and *that* can also be **ergonomic** and *it* kind of feels **a bit different from all the other remote controls**.

In Example 1.2, if a fine-grained opinion analysis system wanted to identify the polarity of each of the opinion expressions, it could potentially look up a sentiment lexicon. This process will be able to determine that the **durable** and **ergonomic** are positive opinions. However, the polarity of **a bit different from other remote controls** is ambiguous – it could be positive or negative depending on the context. Now, if we look at the whole discourse context in which this expression occurs, we observe that the speaker is essentially

[3]Throughout this document the text spans that reveal the opinions are shown in bold, and their targets (what the opinions are about) are shown in italics.

[4]The AMI corpus contains a set of scenario-based face-to-face meetings where a group of four participants have to discuss and design a new TV remote prototype.

speaking about the same thing throughout, the rubbery material. The discourse also tells us that there is a reinforcement – there are no clues to indicate a contrast or juxtaposition. Furthermore, the clear cases of resolved opinions tell us that this reinforcement is in fact of the positive opinion towards the rubbery material. Hence, the opinion polarity interpretation for **a bit different from other remote controls** that is coherent with this discourse is that it is positive. Thus, the discourse context, which in our case comprises of the related opinions, can provide information for disambiguating the polarity of difficult expressions.

In the above example, we saw that the discourse-level treatment is useful for disambiguating fine-grained polarity. Discourse-level analysis can also be useful at a coarser granularity, that is, in determining the overall *stance*. Overall stance classification is an important task in the field of product review mining. In this task, given a web document, the system has to determine whether the writer has an overall positive or negative opinion towards a product. Such systems need to find the relevant expressions of opinions, and from these individual expressions, infer the overall stance. A popular approach here has been to find the individual positive and negative words and aggregate them at the document level to get the overall stance classification. A general simplifying assumption by existing approaches is that all opinions contributing towards a stance are explicitly mentioned in the document. However, this may not always be the case, and in complex cases, a discourse-level treatment will be useful. Let us now consider an Example 1.1 from above (reproduced below) to illustrate this idea. This snippet is from an online debate "iPhone vs. Blackberry, which one is better".

(1.1) *Blackberry* is now for the **senior businessmen market**! The *iPhone* **incarnate the 21st century** whereas *Blackberry* symbolizes an **outdated technology.**

Here, the writer evidently has a pro-iPhone stance. But if the application tries to find the stance regarding the iPhone by looking at the individual opinions, it will only find a single explicit opinion regarding the iPhone. However, looking at the discourse context can reveal more regarding the writer's pro-iPhone stance. Notice that the writer juxtaposes the positive opinions towards the iPhone with negative opinions towards the Blackberry. Additionally, this happens in the context of the debate where the debate topic sets up Blackberry as an alternative to iPhone. In this discourse, the negative opinions towards the Blackberry in

fact reveal more regarding the stance towards the iPhone. That is, the dis-preference for the Blackberry actually is used as a reiteration for a pro-iPhone stance. This example shows that a discourse-level treatment can reveal more about the overall stances.

1.1 RESEARCH SUMMARY

The goal of this research is to **understand and exploit discourse-level relations to improve opinion analysis**.

In order to understand the influence of discourse on opinion analysis, we define and study specific types of discourse-level relations. We then study the impact of these discourse-level relations on improving fine-grained polarity classification and recognition of the overall stance.

1.1.1 Discourse-level relations for opinion analysis

Discourse analysis, in general, finds relations between spans of text. Previous research in NLP has explored relationships between items in the discourse due to coreference, relationships between text spans due to intentional and informational relations (e.g., Rhetorical Structure Theory [Mann and Thompson, 1988], Penn Discourse Treebank [Miltsakaki et al., 2004]), and relations between text spans due to the goals in the discourse (e.g., [Grosz and Sidner, 1986]).

The discourse-level relations in this work are specific to how opinions are expressed in the discourse. Specifically, we explore two sets of discourse-level relations: relations between targets of opinions and relations between opinions themselves[5].

The first set of discourse relations describes relations between the targets of opinions and are called the *target relations*. Two targets are related via a *same* target relation when they refer to essentially the same entity or proposition in the discourse. For instance, in Example 1.2, all the targets are in the *same* relation as they essentially refer to the same entity – the

[5]Note that in this work, relations between opinions are limited to only when their targets are related.

rubbery material.

Targets are related via an *alternative* relation when they refer to mutually exclusive options in the context of the discourse. Example 1.1 illustrates this relation. In this example, iPhone and Blackberry are alternatives due to the discourse context of the debate. That is, only one option can be selected at a time. The notion of alternative relations between items in a discourse is new, but it is not uncommon, and occurs in different genres. Example 1.3 illustrates a scenario from the AMI meeting where an alternative relation exists between the targets.

(1.3) C:: ... shapes **should be** *curved*, so round shapes. **Nothing** *square-like*.

⋮

C:: ... So we **shouldn't have too** *square corners* and that kind of thing.

The targets *curved* shape and *square-like shapes* are mutually exclusive options in the context of this discourse as a given remote control can have only one shape at a time.

The second set of discourse relations explored in this thesis describe relations between opinions and are called the *discourse-level opinion relations*. Two opinions are related via a *reinforcing* relation when they support an overall opinion stance. Examples 1.1, 1.2, and 1.3, each illustrate opinions in reinforcing relationships. In Example 1.1 the opinions reinforce a pro-iPhone stance, in Example 1.2 the opinions reinforce a pro-rubbery material stance, and in Example 1.3 the opinions support a pro-curved shape stance.

On the other hand, opinion pairs that either show ambivalence towards mutually exclusive options, are used for weighing pros and cons of a particular entity, or are, in general, about related targets, but are not used to reinforce a particular stance fall under the category of *non-reinforcing* opinion relations. Examples 1.4 and 1.5 illustrate non reinforcing opinion relations.

(1.4) The *blue remote* is **great**, the *red remote* will be **good** too.

(1.5) The *blue remote* is **cool**, but *it* will be **expensive**.

Specifically, in Example 1.4 we see that the opinions are related in the discourse as their targets, the blue remote and the red remote, are mutually exclusive options (only one of them

can be selected). The opinions are used to convey ambivalence and thus do not reinforce an overall stance. In Example 1.5 the two opinions **cool** and **expensive** are used to weigh the pros and cons of the same target, the blue remote. Hence, they are in a non-reinforcing discourse relation.

Example 1.6 below illustrates an interesting case of non-reinforcing discourse scenario. There are both positive and negative opinions regarding the iPhone indicating non-reinforcement. In fact, this is a special type of non-reinforcing relation characterized by the concessionary construct. The non-reinforcing relations exist between **looks nice** and **can't compare** and between **decent amount** and **can't compare**.

(1.6) While the *iPhone* **looks nice** and does play a **decent amount** of music, *it* **can't compare** in functionality to the BB.

The discourse-level opinion relations in this work are confined to pairs of opinions related via their related targets. Opinion relations automatically exist between polar opinions (opinions with positive or negative polarity) whose targets are related. We represent the opinion relations using structures called *opinion frames*, which are essentially 3-item tuples comprising of the polarities of the two opinions and their target relation. Depending on the value of the contained items, the opinion frames represent a reinforcing or non-reinforcing relation. For instance, the relationships between all pairs of opinions in the Example 1.2 can be represented by the tuple $< positive, positive, same >$. This tuple depicts a reinforcing opinion relation.

1.1.2 Problem statement

The discourse-level relations are exploited for opinion analysis at two different levels. First, they are used to improve fine-grained polarity classification. Here, the information at the discourse layer is used to disambiguate opinions at a finer granularity. Second, the discourse-level relations are also studied and exploited to improve recognition of the overall stance. In this case, the information from the discourse layer is used to improve a recognition task at a coarser granularity.

Thus, this thesis has two high level goals:

Goal-1: Discourse-level Relations for Improving Fine-grained Polarity Recognition

This is the first part of the thesis where we initiate explorations into the discourse-level relations for opinion analysis by creating a linguistic scheme (Chapter 3) and performing corpus annotation and analysis (Chapter 4). Then, we measure the impact of discourse-based methods on fine-grained polarity classification (Chapter 5). Here, we employ the discourse annotations in a global inference framework that augments word-based methods. This part of the thesis also explores modeling choices for augmenting word-based polarity classification with discourse information. We then explore ways in which the manual annotations can be used to automatically learn the discourse relations (Chapter 6).

Goal-2: Discourse-level Relations for Improving Overall Stance Classification

This part of the thesis employs the key concepts about the discourse relations learnt in the first part for performing stance classification. Thus, the information at the discourse-layer is employed to determine the overall stance. Another theme in this part is investigating how our discourse-level relations manifest in real-world data and different domains. Specifically, we carry out experiments in two different types of debates: product debates (Chapter 7) and political and ideological debates (Chapter 8). We do not have manually annotated discourse relations in the debates, and hence we have to find ways to incorporate the ideas behind our discourse relations in a less direct fashion. For product domains, we use web mining and rules to learn and employ the ideas behind the discourse-level relations. In ideological debates, we use user reported stances in a supervised classification framework.

In pursuing the two main goals, this thesis investigates the following high-level hypotheses:

1. *The elements of the discourse-level relations defined in this work can be reliably annotated by trained human annotators.*
2. *The discourse-level relations explored in this work are useful for fine-grained polarity disambiguation over and above word-based methods.*
3. *Automatic systems can be developed for recognizing discourse-level relations better than distribution-based and local information-based baselines.*
4. *The key aspects of the discourse-level relations are useful for stance classification in*

product debates over topic-based baselines and baselines that employ information similar to that used by current product mining systems.

5. *The key aspects of the discourse-level relations are useful for stance classification in political and ideological debates over distribution-based and unigram-based baselines.*

1.1.3 General Approach and Main Results

The two high-level goals described above have the same underlying theme of using a discourse-level treatment for opinion analysis. Our approaches for solving these two problems are, however, distinct.

The first high-level goal encompasses the very first step towards a discourse-level treatment, which is to explore and understand how the discourse-level relations can help opinion analysis. Here, we take a corpus linguistic approach. This includes developing a conceptual representation of the discourse-level phenomena, developing coding schemas and manual annotation instructions for the conceptual representation, producing an annotated corpus, and analyzing the corpus via human annotation studies. Next, empirical studies are conducted to measure the impact of discourse-level relations on fine-grained polarity classification. We then explore linguistic clues and a global inference paradigm for finding the discourse-level relations automatically. We use the AMI meeting corpus for this part of the thesis.

The second high level goal in this thesis is to explore the utility of discourse-level relations to improve overall stance classification. We explore opinion stance classification in two very different domain areas: product debates and ideological debates. For this part of the thesis, we do not use any manually annotated discourse-relations; rather, we capture the key aspects of the discourse relations germane to the task indirectly. In product debates, we explore unsupervised and rule-based methods for incorporating the ideas of the discourse relations for stance classification, while in ideological debates, we incorporate the building blocks of our scheme in a supervised stance classification framework.

We test the first hypothesis, the reliability of the human annotations, by conducting inter-annotator agreement studies. The annotation of our discourse-level relations is a multi-step process, and we measure the reliability at each step. The metrics used for the measure-

ments are similar to those used previously for comparable tasks in NLP. We use retrieval precision metric for measuring the reliability of span annotations (similar to Wiebe et al. [Wiebe et al., 2005]), Cohen's Kappa [Cohen, 1968] for measuring the reliability of labeling tasks such as polarity tagging, and Passonneau's alpha [Passonneau, 2004] for measuring the reliability of identifying discourse-level target relations. Our results for opinion span annotations are comparable to those obtained by Wiebe et al. [Wiebe et al., 2005] for subjectivity span annotations. Similarly, for identifying discourse-level target relations, our reliability results are comparable to Passonneau's results for coreference annotations in monologic texts. Our performance in the labeling tasks such as polarity tagging, type tagging and target link labeling achieve Kappas in reliable range (using the interpretations of the Kappa in [Carletta, 1996, Krippendorff, 2004, Landis and Koch, 1977]).

In order to test the second hypothesis in this work, we employ the discourse-level annotations in a classification framework and measure the impact on fine-grained polarity disambiguation. We first build a system that uses word-based information using state of the art lexical resources. Then, we incorporate the discourse information from the manual annotations using different global inference paradigms. Specifically, we explore an optimization framework (integer linear programming) and a collective classification framework to model discourse relation information. Our results with both these models indicate that discourse-level relation information can significantly improve fine-grained polarity classification over word-based methods.

In order to test the third hypothesis, we explore if and how the discourse-level relations can be learnt in a supervised fashion. Here, we conduct two different sets of experiments, where the first set explores linguistic clues and the second explores global inference paradigms. As a first step, we construct a "local" system that uses only the information in the local context. In this step, the focus is investigation of linguistically motivated features to capture the discourse-level relations. Our results show that, by using discourse-based linguistic features, the presence of discourse-level relations between opinion-bearing sentences can be detected better than distribution-based baseline methods. In another set of experiments, we explore if a global inference method such as collective classification can be used to recognize opinion polarity, discourse-level target relations and discourse-level opinion re-

lations in an iterative, inter-dependent fashion. Thus, in this framework, the output of the opinion polarity system is fed as global information into the discourse-based systems and vice-versa. Our experiments reveal that the global inference paradigm can help to improve the recognition of discourse-based relations over local methods only if some of the global information is non-noisy, that is, if it is provided manually.

In order to test the fourth hypothesis, the utility of discourse-level relations in determining stances in product debates, we use dual-topic online debates discussing which of the given two products is better. While debates are generally popular on the web, debates about specific pairs of products do not provide sufficient data to support supervised approaches. Thus, our stance classifiers for product debates are unsupervised systems. Here, we employ web mining and heuristics to learn a subset of the discourse-level relations that are pertinent for identifying the debate stance. Thus, these experiments also help us to test whether the discourse relations relevant for the task can be learnt in an unsupervised fashion. Specifically, in order to mine the *same* target relations from data on the web we use a PMI-based (Point-wise Mutual Information) measure between the main topics and the targets in debate post. In order to learn the reinforcing relations, we mine weblogs discussing the debate topics and find probabilistic associations. We encode opinion and target information into one single unit called the *opinion-target pair*, and use these units for web mining of the reinforcing relations. We also construct a simple rule-based system, using discourse connective information, to handle special cases of non-reinforcing relations prominent in product debates, namely concessions. We measure the performance of the various systems: the system using only target relations, the system using automatically learnt reinforcing relations and the system using reinforcing and non-reinforcing relations. Our results indicate that the relevant discourse-level relations can be learnt and employed in an unsupervised fashion, and incorporating these can improve stance recognition over baseline methods for debates in the product domain.

The final hypothesis explores the utility of some aspects of the discourse relations for determining stances in political and ideological debates. Here we study if and how the discourse-level relations in this thesis manifest in determining an ideological stance. Examples of such debates are: "Does God Exist", "Creationism" and "Gun Rights". Such

debates are popular on the web. Hence, we have sufficient stance-tagged data for performing supervised stance classification experiments. We employ the building blocks of our discourse relations, the opinion-target pairs, as features in a standard machine learning framework. While the results of our experiments are encouraging, they are not conclusive. We find that the system using sentiment and arguing features in conjunction with the targets (opinion-target pairs) performs better than a simple baseline, and at par with the current state of the art. Our study reveals the complexity of debates in the political and ideological domain. Nevertheless, our analyses show that more insightful information about the stances is captured by our system.

1.2 CONTRIBUTIONS OF THIS WORK

This thesis work pushes opinion analysis into new challenging directions in discourse analysis. The main contribution of this thesis is to establish *discourse-level relations as a useful information source for opinion analysis*. In particular, we show that the discourse-level relation information is useful for tasks at two different granularities: fine-grained polarity recognition and overall stance classification.

This work introduces the idea of two types of discourse-level relations useful for opinion analysis: relations between targets of opinions, and relations between opinion expressions. Discourse-level relations between targets are useful for understanding and capturing the variety of items that people evoke to support their opinions. The *same* target relation defined in this work shares similarities with the previous works in NLP that have looked at relations between items in a discourse due to coreference or bridging descriptions. This work introduces the notion of alternative relation between items – new items that are evoked in the discourse as an alternative to existing items such that, in arguing for these alternatives the speaker, in effect, argues against the original items. This work also introduces the notion of discourse-level relationships between opinions based on whether they reinforce an overall stance. We create a new representation, the opinion frames, for capturing discourse-level opinion relations. The conceptualization of the notion of discourse-level relations and their

representation is presented in [Somasundaran et al., 2008b].

As a part of this dissertation, we adapt the subjectivity annotation scheme developed for news texts [Wiebe et al., 2005] to face to face conversations. In order to account for the change in the genre, we modify the definitions and annotation instructions. We verify whether these annotations can be produced reliably in the meeting genre [Somasundaran et al., 2007a]. Thus, as a consequence of our focus on meeting data, the subjectivity annotations are applied to this new genre.

We extend the original subjectivity annotation scheme to capture the discourse-level relationship between targets. We validate that the annotations can be reliably produced by human annotators. The annotation scheme extensions and validation experiments are presented in [Somasundaran et al., 2008a].

This work is also amongst the first to attempt sentiment and arguing classification in face to face conversations. We explore the usefulness of dialog-based features for opinion type classification. Our results indicate that, in conversations, dialog-level intent is a useful source of information for detecting subjectivity. The experiments for this part have been presented in [Somasundaran et al., 2007a].

This work is one of the first empirical studies to measure the impact of discourse-level relations on fine-grained polarity disambiguation. To this end, this work measures the impact that perfect discourse-based information can have on fine-grained polarity disambiguation. Our results indicate that systems using discourse-level information can achieve substantial and significant improvements over systems using word-based information alone. This thesis also contributes in understanding how to effectively model discourse-based information using global inference paradigms. Our experiments suggest that diverse global modeling frameworks such as optimization and supervised collective classification can be used to effectively capture discourse-level relations to improve fine-grained polarity disambiguation. The experiments and results for this part of the thesis have been presented in [Somasundaran et al., 2009b]

Using our annotated data, we conduct experiments in automatic recognition of the discourse-level relations between pairs of opinion-bearing sentences. The experiments focus on what kind of linguistic and discourse features are useful. These experiments and results

are reported in [Somasundaran et al., 2008b]. We also carry out experiments to determine if fully interdependent joint-inference systems, that is, if systems that use automatic polarity information for discourse-relation classification and vice versa can be created. We show that when global information is non-noisy (at least partially provided by an oracle), the interdependent joint-inference systems can produce overall improvements. These experiments and insights are presented in [Somasundaran et al., 2009a].

For stance classification in product debates, we employ web mining to learn key aspects of our discourse-level relations. The success of our approach shows that ideas behind the discourse-level relations can be learnt and employed in a fully unsupervised fashion. Our approach improves upon a high-precision baseline. We also show that our system that incorporates most of the aspects of our discourse relations is able to perform better than methods used in previous research for product review mining. The results for this part of the work are reported in [Somasundaran and Wiebe, 2009]

Our explorations in classifying stances in political and ideological debates reveal the complexities and challenges involved in ideological stance-taking. Here, we do observe the manifestations of the elements of our discourse-level relations. However, due to the abstract and variable nature of the personal beliefs that embody the ideological stances, we discover that it is difficult to generate general rules. Our first attempts at stance classification using the building blocks of our scheme results in a system that does better than a baseline and as well as the current state of the art that uses unigram information. However, we find that it does capture relatively more insightful details regarding the stance taking than the unigram-based system. Specifically, it not only captures the main issues that are important for the domain, but also captures the concerns, appeals, denials and remonstrations that shape the polarizing stances. In this respect, we believe our explorations sets the groundwork for future research in this field. The experiments and results for this part of the work is reported in [Somasundaran and Wiebe, 2010]

Finally, another byproduct of our research is in the contribution to understanding arguing, a less well explored category in opinion and subjectivity analysis. In this thesis, we annotate this category for meetings. We find that arguing is done in a variety of ways and hence we create extensions to the original definition. We develop resources and explore lin-

guistic features that can help recognize arguing in meetings. Furthermore, arguing is found to be a prominent opinion category in our political and ideological debates. We automatically generate an arguing lexicon for recognizing ideological stances. Our analyses indicate that arguing features are important for recognizing stances in political and ideological debates.

1.3 OUTLINE

The annotation scheme in this thesis builds on the MPQA subjectivity annotation scheme. We briefly give a background of this scheme in Chapter 2. This chapter also provides a background description of the AMI meeting corpus.

In Chapter 3, we present the linguistic scheme for our discourse-level relations. Here, we explain how the scheme can be applied with examples, and also develop intuitions on how these relations will enable us to carry out inter-dependent opinion interpretation. This chapter also discusses how our scheme overlaps and differs from some discourse models used in NLP. Chapter 4 presents the annotation scheme for our discourse-level relations and our human reliability studies for the same. We also explain our efforts to adapt the subjectivity annotations to the meetings, and describe our annotation tool.

Chapter 5 addresses the question of whether the discourse-level relations are useful for fine-grained opinion polarity classification. Here we describe our linguistic clues for polarity classification in meetings, present the development of the global inference paradigms, and report our experimental results using the different models.

In Chapter 6 we address the question of how the discourse-level relations can be learnt from the annotated data. Here, we first present the intuition and results of using linguistically motivated features and then present the formulation and results for the global framework.

Next, we explore stance classification in product debates in Chapter 7. Here, we present our unsupervised systems that attempt to learn the discourse-level relations from the web, describe our rules for employing these relations for stance classification, and report our stance recognition results. In Chapter 8, we describe our work on stance classification in political and ideological debates. Here, we detail the process for constructing the arguing lexicon,

present our experiments and analyze the results.

Finally, in Chapter 9, we summarize the thesis contributions and discuss directions for future work.

2.0 BACKGROUND

The Multi-perspective Question Answering (MPQA) annotation scheme of Wiebe and colleagues and its extension, the fine-grained opinion analysis scheme of Wilson, are the starting point of this work. We describe the aspects of the extended scheme used in this work in Section 2.1.

Our linguistic scheme is developed primarily with the AMI corpus [Carletta et al., 2005] in mind (though we apply the ideas from the scheme to other types of data later on in the thesis). The AMI corpus also provides us with useful tags that we employ as features for our machine learning experiments. We describe the AMI corpus, and its annotations in Section 2.2.

There is a difference between the storage formats in MPQA and AMI. Our process for mapping between these two formats is explained in Section 2.3.

2.1 MPQA ANNOTATION SCHEME AND ITS EXTENSIONS

The fine-grained opinion annotation scheme by Wilson [Wilson, 2007, Wilson and Wiebe, 2005] is based on the MPQA subjectivity annotation scheme from Wiebe [Wiebe et al., 2005, Wiebe, 2002, Wiebe, 1994, Wiebe, 1990], which develops a representation for the idea of private states. A private state is described as the state of an experiencer, holding an attitude, optionally towards a target. The MPQA annotation scheme captures these three main components: the experiencer or source of the private state, the actual expression of the private state, and the target of the private state. Additionally, the private state annotations have attributes such as intensity and polarity. The annotations

are at the sub-sentential level. Sub-sentential analysis of subjectivity involves annotating and recognizing text spans (single or multiple-word expressions) that reveal subjectivity.

Among her extensions, Wilson introduced distinctions between different types of subjectivity, which are called *attitude types*. Wilson identifies the following attitude types:[1]

1. Sentiments: Sentiments are positive and negative emotions, evaluations, and stances. The target of a sentiment is what the sentiment is directed toward.
2. Agreement: Private states in which a person does or does not agree, concede, consent, or in general give assent to something fall into the category of Agreement. The target for this attitude type is what is (or is not) being agreed to.
3. Arguing: Private states in which a person is arguing or expressing a belief about what is true or should be true in his or her view of the world are categorized as Arguing. Arguing attitudes include private states where the source is arguing for or against something. The annotations mark the arguing attitude on the span of text expressing the argument or what the argument is, and mark what the argument is about as the target of the arguing attitude.
4. Intention: Intentions include aims, goals, plans, and other overt expressions of intention.
5. Speculation:Private states in which a person is speculating about what is or is not true, or what may or may not happen, are categorized as Speculation.
6. Other attitude: This is a catch-all category, for the attitudes that do not fall into any one of the above categories.

In this work, we focus on the sentiment and arguing attitude types, and collectively refer to them as *opinions*.

Each attitude annotation has many attributes that capture a wide variety of information. For this work, we are interested in the following elements of the attitude annotation:

id A unique, alphanumeric ID for identifying the attitude annotation.

text anchor A pointer to the span of text that captures the attitude being expressed.

[1]These definitions are from [Wilson, 2008b]. We refer the reader to the original document for more detailed explanations and examples.

attitude type The type of attitude being expressed. The value of this can be positive or negative sentiment, positive or negative arguing, positive or negative intention, positive or negative agreement, speculation and other.

target link This is a list of one or more target IDs. This attribute links attitude annotation to target annotation. Note that this attribute name should not be confused with the "target link" annotations that appear later in this work. In this thesis, "target link" refers exclusively to links between targets.

Wilson's annotations also capture target information for the annotated attitudes. The target annotation consists of the following:

id A unique alphanumeric ID for identifying the target annotation. The ID is used to link the target to the attitude annotation.

text anchor A pointer to the span of text that captures the target.

The attitude annotation scheme is available as MPQA Corpus version 2.0 at http://www.cs.pitt.edu/mpqa. This corpus contains the attitude and target annotations in addition to the original MPQA annotations, and consists of over 344 documents (5957 sentences).

2.2 AMI CORPUS

The AMI Meeting Corpus consists of 100 hours of meeting recordings. The recordings use a range of signals synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. The meetings are recorded in English.

In this work, we use the scenario-based meetings from the AMI corpus. Scenario-based meetings consist of four participants in a goal-oriented meeting, where the goal is to design a new TV remote prototype for an electronics company. The participants play different roles in a fictitious design team that takes a new project from kick-off to completion over the course of a day. The day starts with training for the participants about what is involved

in the roles they have been assigned (industrial designer, interface designer, marketing, or project manager) and then contains four meetings, plus individual work to prepare for them and to report on what happened.

The AMI Meeting Corpus includes high quality, manually produced orthographic transcription for each individual speaker, including word-level timings that have been derived by using a speech recognizer in forced alignment mode. The AMI meetings provides automatic segmentation of the meetings into segments (segments are similar to utterances). It also contains a wide range of other annotations, not just for linguistic phenomena but also detailing behaviors in other modalities. These include dialog acts; topic segmentation; extractive and abstractive summaries; named entities; the types of head gesture, hand gesture, and gaze direction that are most related to communicative intention; movement around the room; emotional state; and where heads are located on the video frames. We are particularly interested in the Dialog Act (DA) annotations. We will explain these in more detail in the next section.

2.2.1 Dialog Act Annotations

Dialog Act (DA) annotations code speaker intentions, and segment the transcripts into separate dialog acts classified as follows: acts about the information exchange (Inform, Elicit-Inform), acts about possible actions (Suggest, Offer, Elicit-Offer-or-Suggestion), comments on previous discussion (Access, Comment-About-Understanding, Elicit-Access, Eilicit-Comment-About-Understanding), social acts (Be-positive, Be-negative), and a special set of tags for labeling non-intentional acts e.g. backchannels, such as "um" and "uh-huh", and attempts to maintain the floor. The latter category plus an additional bucket class for all other intentional acts allow for complete segmentations, with no part of the transcript left unmarked.

Each of these Dialog Acts is described as follows:[2]

- Inform: The Inform act is used by a speaker to give information.

[2]These definitions are reproduced from http://mmm.idiap.ch/private/ami/annotation/dialogue_acts_manual_1.0.pdf. Please refer to this document for more complete explanations and examples.

- Elicit-Inform: The Elicit-Inform act is used by a speaker to request that someone else give some information.
- Suggest: In a Suggest, the speaker expresses an intention relating to the actions of another individual, the group as a whole, or a group in the wider environment.
- Offer: In an Offer, the speaker expresses an intention relating to his or her own actions.
- Elicit-Offer-Or-Suggestion: In a Elicit-Offer-Or-Suggestion, the speaker expresses a desire for someone to make an offer or suggestion
- Assess: An Assess is any comment that expresses an evaluation, however tentative or incomplete, of something that the group is discussing, where the something could be another dialog act or something apparent from the working environment, like slides or, in the remote control design trials, the playdough remote control mock-up. There are many different kinds of assessment; they include, among other things, accepting an offer, expressing agreement/disagreement or any opinion about some information that's been given, expressing uncertainty as to whether a suggestion is a good idea or not, evaluating actions by members of the group, such as drawings. Assessments themselves can be assessed in further acts, and if the thing being assessed is a dialog act, it doesn't have to be from a different speaker, since people can assess their own contributions to the discussion.
- Comment-About-Understanding: Comment-About-Understanding is for the very specific case of commenting on a previous dialog act where the speaker indicates something about whether they heard or understood what a previous speaker has said, without doing anything more substantive.
- Elicit-Assessment: In an Elicit-Assessment, the speaker attempts to elicit an assessment (or assessments) about what has been said or done so far.
- Elicit-Comment-About-Understanding: In an Elicit-Comment-About-Understanding, the speaker attempts to elicit a comment about whether or not what has been said or done so far has been understood, without further asking for assessment of anything in the discussion.
- Be-Positive: Be-Positive includes any social acts that are intended to make an individual or the group happier. For example, little acts of politeness like greeting one another or

saying "please", "sorry", and "thank you".

- Be-Negative: Be-Negative includes any social acts that express negative feelings towards an individual or the group.
- Other: The Other class is a bucket for all proper dialog acts - where the speaker is conveying an intention that don't fit any of the other classes.

2.3 CONVERTING FROM THE AMI FORMAT TO THE MPQA FORMAT

The AMI annotations of our interest are stored based on words and/or time information. Specifically, the AMI corpus provides meeting transcriptions in the form of word information and the start and end times of the words in the meeting. Similarly, segment information is also time based. Each of the scenario-based AMI meetings have four participants, and the words uttered by each participant (and the segments containing them) is provided in a separate file.

The MPQA annotation scheme, on the other hand, is byte-span based. As our annotations are based on and intended as an extension to the MPQA annotation scheme, we convert from the AMI format to the MPQA format. This is done for each AMI meeting as follows:

- The words contained in each segment are found using the start and end times of the segments and the words. A word belongs to a segment if its start time is greater than or equal to the start time of the segment, and its end time is less than or equal to the end time of the segment.
- All the words in each segment are sorted according to their start time
- All the segments for all the participants (the participant IDs are retained for the segments) are merged into a single file and sorted according to their start time. This gives us a single file containing all the utterances for each meeting.

This is the resultant MPQA document that we use for our annotations and experiments. We also store an information "bridge" file, containing the byte spans for each word from

the MPQA document and the AMI start and stop times for that word. This file is used to map information from the AMI format to the MPQA format and vice versa. The following line shows an example of the information stored in this file.

```
34,39 string AMI_Word token=Right
stTime=53.56 endTime=53.96 id=ES2002a.B.words2
isPunc=false isVocal=false vocaltype=null isGap=false isDisfMarker=false
```

The byte span (34,39) corresponds to the byte span of the word in the MPQA document. The actual word token is "Right". Its AMI ID is ES2002a.B.words2 (word 2, from speaker B, in meeting ES2002a), and its start time in the AMI meeting is 53.56, and its end time is 53.96. The remaining information, namely that it is not a punctuation,vocalization, gap or disfluency marker is the additional information provided by AMI for the word.

AMI Dialog Act (DA) annotations are encoded with respect to the words that each DA spans. We use the bridge file to map DA information (DA-based meeting segmentation and DA tags) to our annotations.

3.0 LINGUISTIC SCHEME FOR DISCOURSE-LEVEL RELATIONS

Our goal in this chapter is to **create a linguistic scheme to represent our ideas of discourse-level opinion relations**.

Data annotation and human reliability studies are carried out for this scheme in Chapter 4. It is used directly for experiments in Chapter 5, where we employ the annotations from the scheme to improve polarity classification. It inspires our approach to stance classification in Chapters 7 and 8. For product debates, we use web mining and heuristics to learn and apply the elements of the scheme useful for the task whereas for political debates we use the user-tagged data and attempt to learn the elements of the scheme in a supervised fashion.

As a first step in the explorations of our problem space, we will develop the intuition behind the discourse-level relations. We will then present and explain the definitions and the representation of our sceheme (Section 3.1). Via examples, we show how the scheme can be applied on the data (Section 3.2). Our scheme is developed with interdependent opinion disambiguation in mind, which we explain in Section 3.3. We will have a detailed discussion of how the scheme relates to some major NLP discourse models, where we find overlaps and differences in Section 3.4. Other related work is discussed in Section 3.5 and finally we summarize and discuss future directions in Section 3.6.

3.1 SCHEME

The linguistic scheme for discourse-level relations is based on the idea that *opinion expressions are related in the discourse via the relation between their targets and whether/ how the opinions contribute to an overall stance*. Conceptually, we have two types of discourse-level

relationships: the discourse-level target relations and the discourse-level opinion relations.

Targets of opinions may be either unrelated, or related via a *same* or an *alternative* relation. Target relations indicate if opinions in the discourse are about related entities or propositions. Let us look at the target relations in the following examples which we saw previously (reproduced below).

(1.1) *Blackberry* is now for the **senior businessmen market**! The *iPhone* **incarnate the 21st century** whereas *Blackberry* symbolizes an **outdated technology**.

(1.2) D:: ... this kind of rubbery material, *it's* a **bit more bouncy**, like you said they get chucked around a lot. A **bit more durable** and *that* can also be **ergonomic** and *it* kind of feels **a bit different from all the other remote controls**.

(1.3) C:: ... shapes **should be** *curved*, so round shapes. **Nothing** *square-like*.

⋮

C:: ... So we **shouldn't have too** *square corners* and that kind of thing.

(1.6) While the *iPhone* **looks nice** and does play a **decent amount** of music, *it* **can't compare** in functionality to the BB.

Notice that, in Example 1.2, all the opinions are essentially about the same thing, the rubbery material. The first target *it's* refers to the rubbery material directly. The target of the opinion **bit more durable** is an ellipsis that refers to the rubbery material. The *that* refers to the property of the rubbery material of being durable, and finally the *it* again directly refers to the rubbery material. Similarly, the targets *iPhone* and *it* in Example 1.6 refer to the same object, the iPhone.

On the other hand, the targets *curved* and *square-like* in Example 1.3 are in a very different kind of relationship. In the context of this discourse, curved shapes and square shapes are alternatives or mutually exclusive options (the TV remote can have only one shape at a time). A similar alternative relation is seen between the targets *Blackberry* and *iPhone* in Example 1.1. Here the context of the debate (iPhone vs. Blackberry) sets them up as alternatives.

With these observations in mind, we can now formally define the two target relations.

Same target relation: The *same* relation holds between targets that refer to the same

entity, property, or proposition. Observing the relations marked by annotators, we found that *same* covers not only identity, but also part-whole, synonymy, generalization, specialization, entity-attribute, instantiation, cause-effect, epithets and implicit background topic, i.e., relations that have been studied by many researchers in the context of anaphora and co-reference (e.g., [Clark, 1975, Vieira and Poesio, 2000, Mueller and Strube, 2001]). Actually, *same* relations holding between entities often involve co-reference (where co-reference is broadly conceived to include relations such as part-whole listed above). However, there are no morpho-syntactic constraints on what targets may be. Thus, *same* relations may also hold between adjective phrases, verb phrases, and clauses.

Alternative target relation: The *alternative* relation holds between targets that are related by virtue of being opposing (mutually exclusive) options in the context of the discourse. For example, in the domain of TV remote controls, the set of all shapes are alternatives to one another, since a remote control may have only one shape at a time. In such scenarios, a positive opinion regarding one choice may imply a negative opinion toward competing choices, and vice versa. Objects appear as alternatives via world and domain knowledge (for example, shapes of a remote) or the context of the discourse (for example, Hillary Clinton and Barack Obama are alternatives in discussions of the democratic primaries, but not in discussions of the general election). Alternatives can also be established explicitly by a discourse participant (for example, the following utterance "which car do we want to buy, the Lexus or the Acura" explicitly sets up the two cars as alternatives in the discourse).

In order to represent the relationships between opinions, we define a structure called *opinion frames*, which consist of opinion pairs that are related by virtue of having united or opposed target relations. Opinion frames are essentially 3-item tuples consisting of the two opinions and their target relation. Once we have the target relations for a given opinion pair, we can construct the opinion frame for them using the target relation and their individual opinion information. Depending on the value of its components, the opinion frame denotes whether the opinions are in a *reinforcing* or *non-reinforcing* discourse relation.

Now, let us look at the discourse-level relationships between the opinion expressions in our examples. In Example 1.2, we see that all the opinion expressions are positive, they

are all towards targets participating in *same* pairwise relationships. The first two opinions (**bit more bouncy** and **bit more durable**) are direct positive evaluations of the rubbery material. The third opinion (**ergonomic**) positively evaluates the durable property of the rubbery material and finally, the opinion **a bit different from all the other remote controls** is a positive evaluation of the rubbery material. Here all the opinions are used to gather support or reinforce an overall pro-rubbery material stance, and hence they are said to be in a reinforcing relationship.

Next, if we analyze the opinions **should be** and **Nothing** in the Example 1.3, we see that they are of opposite polarity. Also, they have a different type of target relation – the alternative relation. However, they are still unified in supporting an overall stance: a pro-curved shape stance. In particular, the first opinion is a direct explicit support of the stance while the second opinion provides indirect support for the stance via a negative opinion towards the alternative. The third opinion in the passage (**shouldn't have too**) also supports the overall pro-curved shape stance via a negative opinion towards the mutually exclusive option. Hence, in this discourse too, all the opinions are said to reinforce one another.

On the other hand, the positive and negative opinions in Example 1.6 are in a different kind of relationship. The positive opinion **decent amount** and the negative opinion **can't compare** are both towards the iPhone. The opposition in the polarities towards the same target indicates that these opinions are not reiterating a support for a pro-iPhone stance.

With these observations in mind, we can now formally define the opinion relations.

Reinforcing opinion relation: The reinforcing relation exists between opinions when they contribute to the same overall stance.

Non-reinforcing opinion relation: The non-reinforcing relation exists between opinions that show ambivalence, concession or other conditions where the opinions are related (via a target relation), but do not support an overall stance.

It is important to remember that we do not create the reinforcing/non-reinforcing relations directly between the opinions. Instead, we construct opinion frames and these represent the discourse-level relations between the opinions. Depending on the value of the components, opinion frames can denote reinforcing or non-reinforcing relations. Table 3.1 lists the

Discourse-level opinion relations	Opinion Frames using only polarity	OPinion frames using polarity and opinion type information A= Arguing , S= Sentiment
Reinforcing	$< +, +, same >$	$< A+, A+, same > < A+, S+, same >$ $< S+, A+, same > < S+, S+, same >$
	$< -, -, same >$	$< A-, A-, same > < A-, S-, same >$ $< S-, A-, same > < S-, S-, same >$
	$< +, -, alt >$	$< A+, A-, alt > < A+, S-, alt >$ $< S+, A-, alt > < S+, S-, alt >$
	$< -, +, alt >$	$< A-, A+, alt > < A-, S+, alt >$ $< S-, A+, alt > < S-, S+, alt >$
Non-reinforcing	$< +, -, same >$	$< A+, A-, same > < A+, S-, same >$ $< S+, A-, same > < S+, S-, same >$
	$< -, +, same >$	$< A-, A+, same > < A-, S+, same >$ $< S-, A+, same > < S-, S+, same >$
	$< +, +, alt >$	$< A+, A+, alt > < A+, S+, alt >$ $< S+, A+, alt > < S+, S+, alt >$
	$< -, -, alt >$	$< A-, A-, alt > < A-, S-, alt >$ $< S-, A-, alt > < S-, S-, alt >$

Table 1: Discourse-level opinion relations and the opinion frames that represent them

different discourse-level opinion relations and the frames that represent these relations. If we use only polarity information of the related opinions, we get the frames listed in Column 2 of Table 3.1. On the other hand, if we distinguish the opinion type along with the polarity, we get the opinion frames listed in Column 3.

Each opinion frame from the reinforcing category in Table 3.1 depicts a scenario where the opinions, and their target relations reinforce an overall stance. For example, the frame $< +, +, same >$ represents a discourse where the speaker is repeating his positive opinion about a target and frame $< -, -, same >$ denotes a discourse where the speaker is reinforcing his dis-preference. Opinion frame $< +, -, alt >$ describes a discourse scenario where the speaker is employing opinions of opposite polarities towards alternatives to reinforce an overall stance. The non-reinforcing frame $< +, -, same >$ characterizes a non-reinforcing scenario where the speaker is ambivalent, and perhaps weighing the pros and cons of an item. Likewise, the frame $< -, -, alt >$ depicts another non-reinforcing scenario, where the speaker dislikes both alternatives, and hence the related opinions do not reinforce a stance towards either one of the options.

3.2 EXAMPLES

Let us now illustrate the application of our linguistic scheme with some examples from the AMI meeting corpus. Figure 1 illustrates the opinion and target spans, the target relations, the resulting opinion frames and the corresponding opinion relations for the Example 1.2. In the figure, the speaker's positive sentiment regarding the rubbery material is apparent from the text spans **bit more bouncy** (+), **bit more durable** (+), **ergonomic** (+) and **a bit different from all the other remote controls** (+). As shown, the targets of these opinions (*it's* (t1), *that* (t3), and *it* (t4)) are related by the *same* relation. The ellipsis occurs with **bit more durable**. Target t2 represents the (implicit) target of that opinion, and t2 has a *same* relation to t1, the target of the **bit more bouncy** opinion. The opinion frames occurring throughout this passage are all $< +, +, same >$ denoting that both the opinion components are positive with a *same* relation between their targets. One frame

occurs between O1 and O2, another between O3 and O4, and so on. Each of these opinion frames denote a reinforcing relation.

Similarly Figure 2 illustrates the different elements of the conceptual scheme for Example 1.3. In this example, we see one positive arguing (+) **should be**, and two negative arguing (−) **Nothing** and **shouldn't have too**. The targets of these opinions are *curved*, *square-like* and *square corners* respectively. Targets t1 and t2 are in an *alternative* target relation. The targets t2 and t3 essentially refer to the same concept – square shaped remotes. Hence, t2 and t3 are in a *same* target relation. As t3 is the same as t2, it is also an alternative to t1, as indicated in the figure.

Figure 3 shows the application of the scheme on Example 1.1. Note that this example is from a product debate ("iPhone vs. Balckberry"). Here the opinions are all of the sentiment type. Specifically, there is a negative sentiment O1 (**senior businessmen market**) towards the *Blackberry* (t1), a positive opinion O2 (**incarnate the 21st century**) towards the *iPhone* (t2) and finally a negative opinion O3 (**outdated technology**), towards the *Blackberry* (t3). The debate context sets up iPhone and Blackberry as alternatives in this discourse. Thus, targets t1-t2 and t2-t3 have alternative target relations. In spite of all the variety of the opinion polarities and the target relations, all the opinion frames in this passage are of reinforcing type – they reinforce a pro-iPhone stance

Figure 4 illustrates the discourse-level relations for another example from the iPhone vs. Blackberry debate. All the opinions in this sentence are towards it same target – the iPhone. The first two opinions O1 and O2 (**looks nice** and **decent amount**) are positive while the third opinion O3 (**can't compare**) is negative. As the speaker has conflicting opinions regarding the same object, we see some non-reinforcing frames in this example. Specifically, the opinions O1-O3 and O2-O3 are in non-reinforcing discourse-level relations. Both these relations are represented by the $< +, -, same >$ opinion frame.

D:: ... this kind of rubbery material, *it's* a **bit more bouncy**,like you said they get chucked around a lot. A **bit more durable** and *that* can also be **ergonomic** and *it* kind of feels**a bit different from all the other remote controls**.

Opinion	Polarity	Text Span	Target	Text span
O1	+	**bit more bouncy**	t1	*it's*
O2	+	**bit more durable**	t2	ellipsis
O3	+	**ergonomic**	t3	*that*
O4	+	**a bit different from all the other remote**	t4	*it*

Target - target	Target Relation Type
t1 - t2	same (ellipsis)
t3 - t4	same (identity)
t1 - t3	same (identity)

Opinion-Opinion	Opinion Frame	Opinion Relation Type
O1 - O2	$< +, +, same >$	reinforcing
O1 - O3	$< +, +, same >$	reinforcing
O3 - O4	$< +, +, same >$	reinforcing

Figure 1: Discourse-level relations for Example 1.2

C:: ... shapes **should be** *curved*, so round shapes. **Nothing** *square-like*.

⋮

C:: ... So we **shouldn't have too** *square corners* and that kind of thing.

Opinion	Polarity	Text Span	Target	Text span
O1	+	**should be**	t1	*curved*
O2	−	**Nothing**	t2	*square-like*
O3	−	**shouldn't have too**	t3	*square corners*

Target - target	Target Relation Type
t1 - t2	alternatives
t2 - t3	same (same concept)
t1 - t3	alternatives

Opinion-Opinion	Opinion Frame	Opinion Relation Type
O1 - O2	$< +, -, alt >$	reinforcing
O2 - O3	$< -, -, same >$	reinforcing
O1 - O3	$< +, -, alt >$	reinforcing

Figure 2: Discourse-level relations for Example 1.3

Blackberry is now for the **senior businessmen market**! The *iPhone* **incarnate the 21st century** whereas *Blackberry* symbolizes an **outdated technology**.

Opinion	Polarity	Text Span	Target	Text span
O1	−	**senior businessmen market**	t1	*Blackberry*
O2	+	**incarnate the 21st century**	t2	*iPhone*
O3	−	**outdated technology**	t3	*Blackberry*

Target - target	Target Relation Type
t1 - t2	alternatives (debate context)
t2 - t3	alternatives (debate context)
t1 - t3	same (identity)

Opinion-Opinion	Opinion Frame	Opinion Relation
O1 - O2	$< -, +, alt >$	reinforcing
O2 - O3	$< +, -, alt >$	reinforcing
O1 - O3	$< -, -, same >$	reinforcing

Figure 3: Discourse-level relations for Example 1.1

While the *iPhone* **looks nice** and does play a **decent amount** of music, *it* **can't compare** in functionality to the BB.

Opinion	Polarity	Text Span	Target	Text span
O1	+	**looks nice**	t1	*iPhone*
O2	+	**decent amount**	t2	*iPhone*
O3	−	**can't compare**	t3	*it*

Target - target	Target Relation Type
t1 - t2	same
t2 - t3	same (identity)
t1 - t3	same (identity)

Opinion-Opinion	Opinion Frame	Opinion Relation Type
O1 - O2	$< +, +, same >$	reinforcing
O2 - O3	$< +, -, same >$	non-reinforcing
O1 - O3	$< +, -, same >$	non-reinforcing

Figure 4: Discourse-level relations illustrating non-reinforcing relations

3.3 INTERDEPENDENT INTERPRETATIONS

The interdependent nature of opinions and discourse relations can help in their mutual disambiguation.

For instance, we see that in Example 1.2, when some aspect of an opinion such as the polarity is unclear, the discourse-level relations can help to disambiguate the polarity of the instance by bringing in the information of the related opinions and the discourse context. Specifically, here we see that out of context, the polarity of **a bit different from other remotes** is unclear. However, the polarities of two of the opinions are clear (**durable** and **ergonomic**). There is evidence in this passage of discourse continuity and *same* relations, such as the pronouns, the lack of contrastive cue phrases, and so on. This evidence suggests that the speaker expresses similar opinions throughout the passage, making the opinion frame $< +, +, same >$ more likely throughout. Recognizing the frames would resolve the polarity ambiguity of **a bit different from other remotes**. This example is a case where the discourse-level relations are clear, and the resolved opinions in the clear discourse context helps disambiguating a difficult expression.

The direction of inference may also be turned around. That is, if there are clear opinion expressions in the discourse, these can help in the disambiguation of discourse relations. In the following example (Example 3.1), the positive sentiment (+) towards *this* and the positive arguing (+) for *it* are clear. These two individual opinions can be related by a *same*/*alternative* target relation, be unrelated, or have some other relation not covered by our scheme (in which case we would not have a relation between them). There is evidence in the discourse that makes one interpretation more likely than others. The "so" indicates that the two clauses are highly likely to be related by a Cause discourse relation (PDTB). This information confirms a discourse continuity, as well as makes a reinforcing scenario likely, which makes the reinforcing frame $< +, +, same >$ highly probable. This increase in likelihood will in turn help us to infer that *this* and *it* co-refer (and are thus in a *same* target relation).

(3.1) B :: ... and *this* will **definitely enhance our market sales**, **so we should** take *it* into consideration also.

3.4 DISCUSSION

In this section we discuss examples and illustrate how the discourse-level relations defined in this thesis compare to some discourse relations commonly used in NLP. Specifically, we compare our discourse-level target relations to coreference relations and our discourse-level opinion relations to discourse relations from the Penn Discourse Treebank. We also discuss how our relations based on opinions compare with dialog structure annotations in the AMI corpus.

3.4.1 Coreference

In our definitions for target relations, we mentioned that the *alternative* relations are novel in this work. However, the *same* target relations overlap with coreference relations.

Consider Example 1.2 (see Figure 1). Observe that the *same* target relation between t1 and t4 overlaps with the anaphoric coreference relation existing in the discourse. Target t2 is an ellipsis, and t3 (*that*) is a deixis that refers to the rubbery material's property of being durable. Thus, in this particular example, there is a nice overlap between the coreference phenomena and the *same* target relation.

However, targets are not confined to any particular part of speech category – they may be noun phrases, adjectives, verbs or adverbs. They are simply defined as *what* the opinion is about. Consequently, the targets of opinions can be entities, propositions, events, situations, etc. They may be single words, or multiple-word expressions. As a result, target relations could exist between words/phrases of different part of speech, different lengths and even different semantic types.

Example 1.3 (see Figure 2) illustrates a scenario where target relations exist between words from different part of speech categories. In this example, *curved* and *square-like* are adjectives. The target relation t1-t2 is between adjectives. The *square corners* is a noun phrase. Thus, the relation t2-t3 is between an adjective and a noun phrase.

(3.2) C :: Oh in that case **you can you always** *hook up with someone* who is providing that and you know, you sell their product as well as your product with them ...

D :: Yeah, but **we want** to *design a new one.*

Example 3.2 illustrates how the targets of opinions can be multiple-word expressions. Here, speaker C argues for partnering with some company that makes the item under discussion. Speaker D is opposed to this, and argues for designing the item themselves. The options of collaboration and designing it themselves are mutually exclusive situations. These targets are linked via an alternative target relation.

Hence, even though there are some situations where target relations and coreference relations overlap, our target relations are not redundant with coreference relations.

3.4.2 Discourse Relations

Rhetorical Structure Theory (RST) [Mann and Thompson, 1988], the Penn Discourse Treebank (PDTB) [Miltsakaki et al., 2004], Groz and Sidner's theory of attentional states [Grosz and Sidner, 1986], and Hobbs theory of discourse coherence [Hobbs, 1979, Hobbs et al., 1993] are some of the popular models in discourse analysis. Discourse theories define relations between text spans in the discourse. The main difference between these and our discourse-level opinion relations is that our discourse-level relations are specific – these are relations only between opinions spans (and in this work, the opinion relations are additionally confined to those opinions in the discourse that are related via target relations).

RST relations and our discourse-level opinion relations further differ due to the fact that we do not have a concept of nucleus and satellite. Additionally, there is difference in the discourse structure resulting from the relations. RST considers the discourse to have a tree structure. We do not attempt to impose any structure on the discourse. Our pair-wise relationships eventually result in a discourse-level graph. Similarly, Groz and Sidner's theory pays attention to the structure of discourse, our discourse-level relations' definition does not have this focus.

The Penn Discourse Treebank scheme annotates discourse connectives (both explicit and implicit occurrences) and annotates clausal arguments to the connectives. This scheme relates pairs of text spans by discourse relations such as Contrast, Elaboration, List, Con-

junction, Temporal relations, etc. We compared the relations produced by our scheme to the ones produced by the PDTB annotations. We found that there are situations where the PDTB annotations and our annotations overlap. Figure 5 illustrates such a situation.

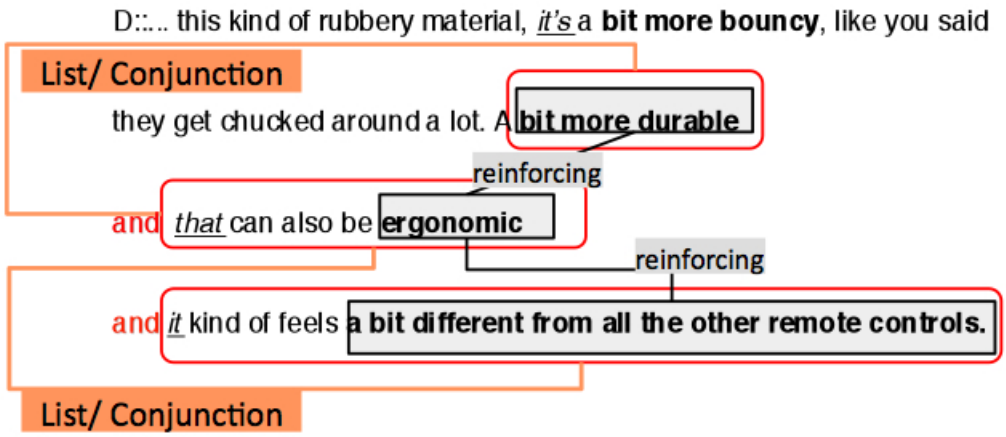

Figure 5: **Correspondence between PDTB and Discourse-level opinion relations**

In Figure 5, we see that there is an explicit discourse connective "and" between the text spans "A bit more durable", "that can also be ergonomic" and "it kind of feels a bit different from other remote controls". Consequently, the text spans would be related via the List or Conjunction PDTB discourse relations. Our opinion relations also correspond with approximately the same text spans (shown in grey boxes in the figure). All the opinion relations in this passage are reinforcing. This example illustrates the case in which PDTB relations nicely correspond to opinion frame relations. The opinion frames flesh out the discourse relations: we have lists specifically of positive sentiments toward related objects.

However, opinion relations and discourse-relation schemes are not redundant. Consider the following three passages.

(3.3) D:: ... I draw for you this *schema* that can be maybe **too technical** for you but is **very important** for me ....

[Non-reinforcing opinion relation (< −, +same >); Contrast discourse relation]

(3.4) D:: **not too** *edgy* and *like a box*, **more** kind of *hand-held*

[Reinforcing opinion relation (< −, +, alt >); Contrast discourse relation]

(3.5) ... they **want** something that's *easier to use* straight away, *more intuitive* perhaps.
[Reinforcing opinion relation (< +, +same >); no discourse relation]

In both Examples 3.3 and 3.4, the PDTB discourse relation between the text spans containing the opinions is Contrast. In particular, "too technical for you" is contrasted with "very important for me" in Example 3.3 and "not too edgy and like a box" is contrasted with "more kind of hand-held" in Example 3.4. However, the opinion frame in Example 3.3 is $< -, +, same >$, which is a non-reinforcing frame, while the opinion frame in Example 3.4 is $< -, +, alt >$, which is a reinforcing frame. In Example 3.5, the opinion relation holds between targets within a subordinated clause (*easier to use* and *more intuitive* are the two targets of **want**); most discourse theories (such as RST, Hobbs theory of coherence and PDTB) don't predict any discourse relation in this situation.

Generally speaking, we find that there are not one-to-one mappings between opinion frames and the relations of popular discourse theories.

3.4.3 Dialog Acts

Dialog Acts are focused on interpersonal exchanges and discourse functions, while our opinion categories are focused on participants' private states, usually towards objects (which incidentally may be other participants).

We found that the scheme for dialog structure, Dialog Acts (DA) and Adjacency Pairs, defined in AMI meetings [Carletta et al., 2005] and our opinion categories are complementary rather than interchangeable. The scheme for dialog structure focus on the speaker intent with respect to the the ongoing conversation. That is, whether the speaker is asking a question, answering a question, conveying information, being positive (conversational niceties), etc.

Specifically, in the AMI corpus, we found that the Access Dialog Act and the sentiment opinion type overlap sometimes, but not always. The example below illustrates this. Here the text span containing **pretty good** is an assessment (Assess) DA, illustrating an overlap between the positive sentiment towards the object and the positive assessment at the dialog level. However, the first Assess DA annotation ("Well yeah") in the example is an assessment

purely at the dialog level and has no corresponding opinion annotation.

(3.6) < *Assess* >Well yeah, < */Assess* > < *Assess* >I mean compared to most remote controls you see *that's* **pretty good** < */Assess* >

Our opinions also occur with other DA types as seen in the example below. The utterance is labeled as an Inform DA as it functions to inform the participants of the roomy interior of the remote control. Orthogonally, it is tagged as a positive sentiment (**very easy to use**) and positive arguing (**'Cause**).

(3.7) Okay, so when you have a *lot of room inside*. So you can make it **very easy to use**. **'Cause** you can write a lot of comments besides it.

3.5 RELATED WORK

Polanyi and Zaenen [Polanyi and Zaenen, 2006], in their discussion on contextual valence shifters, have also observed the phenomena described in this work – namely that a central topic may be divided into subtopics in order to perform evaluations, and that discourse structure can influence the overall interpretation of valence. However, they do not propose a definite linguistic scheme. In another parallel research, Stoyanov et al. [Stoyanov and Cardie, 2008a] annotate co-referent opinions, that is, opinions referring to the same overall topic. The relation defined in Stoyanov et al. is closer to topical coreference. Our target relations are more detailed – we have different types of relations and coreference is a subtype within our *same* target relation.

Incorporating global information in the form of discourse information is gaining importance for opinion analysis. Asher et al. [Asher et al., 2008], in a parallel work, advocate discourse-level analysis in order to get a deeper understanding of contextual polarity and the strength of opinions. Their scheme is based on RST relations, and tries to build a tree structure in the discourse. In contrast, our scheme focuses on pair-wise relations directly between opinion expressions. Snyder and Barzilay [Snyder and Barzilay, 2007] combine an agreement model based on contrastive RST relations with a local aspect model to make a

more informed overall decision for sentiment classification. Theirs is not a linguistic scheme, rather a machine learning approach based on agreement between low-level tags.

There are also other schemes for dialog related phenomena. The most closely related work includes the dialog-related schemes for various available corpora of conversation (for example, SWBD DAMSL [Jurafsky et al., 1997], ICSI MRDA [Dhillon et al., 2003] and ISL [Burger et al., 2002, Burger and Sloane, 2004]). The SWBD DAMSL annotation scheme over the Switchboard telephonic conversation corpus labels shallow discourse structures. SWBD DAMSL had a label "sv" for opinions. However, due to poor inter-annotator agreement, the authors discarded these annotations. The ICSI MRDA annotation scheme adopts the SWBD DAMSL scheme, but does not distinguish between opinionated and objective statements. The ISL meeting corpus is annotated with dialog acts and discourse moves such as initiation and response, which in turn consist of dialog tags such as query, align, and statement. Their statement dialog category would not only include sentiment and arguing opinion categories, but it would also include objective statements and other types of subjectivity.

Most other work in multiparty conversation has focused on exchange structures and discourse functional units like common grounding [Nakatani and Traum, 1998]. In common grounding research, the focus is on whether the participants of the discourse are able to understand each other, and not their opinions towards the content of the discourse. Other schemes like the one proposed by [Flammia and Zue, 1997] focus on information seeking and question answering exchanges where one participant is purely seeking information, while the other is providing it. In general, most dialog-based schemes are orthogonal to ours. Our scheme focuses exclusively on the opinion content and is not exclusive to dialogs.

Another related work finds participants' positions regarding issues via argument diagramming [Rienks et al., 2005]. This approach, based on the IBIS system [Kunz and Rittel, 1970], divides a discourse into issues, and finds lines of deliberated arguments. However they do not distinguish between subjective and objective contributions towards the meeting. Generally, research in the filed of argumentation (e.g. [Toulmin, 1969, Cohen, 1987]) deals with claims and warrants. Here the focus is on the presentation of facts that support claims, while this thesis focuses on subjective arguing.

3.6 SUMMARY AND FUTURE WORK

In this chapter, we described the linguistic scheme that forms the foundations of our work on discourse-level relations for opinion analysis. Specifically, we define two types of discourse-level relations, the target relations and the opinion relations. We define a tuple-based representation, the opinion frames, for representing the relations between opinions. We show, via illustrated examples, the relations that are created when the scheme is applied and discuss the interdependent interpretation between opinion expressions and discourse-level relations. We also discuss how and where our scheme overlaps and differs from other discourse analysis schemes.

Our linguistic scheme presented in this chapter is just a first step in exploring the problem space of discourse-level relations between opinions. Both opinion relations and target relations can be extended more comprehensively. We discuss two potential extensions, one for the opinion relations and one for the target relations below.

In this work, the discourse-level relations between opinions are confined to opinions with related targets. This requirement was useful as a first step, as it established clear and tangible relationships. However, it is possible for opinions to relate to one another without having a target relation. For instance, consider the following example:

(3.8) B :: It's **no matter**.
D :: Okay , it's y yeah ..
B :: **No problem**.

In Example 3.8, the two opinions **no matter** and **No problem** are related – they reinforce an opinion stance that something (that is not evident is the immediate discourse context) does not matter. However, the current scheme does not recognize this relation between the opinions due to the absence of explicit targets.

Also, even when opinions do have targets, it is not always possible to relate the targets via *same* and *alternative* relations. The following example illustrates this situation.

(3.9) A :: **But we can't really** *design for something that hasn't been invented yet.*
C :: Ah *it's it's* **coming up, actually**. *The personal video recorder and all those*

things **it is coming up.**

In Example 3.9, we see that the opinions of speakers A and C are related in the discourse. Speaker A is arguing that they can not create the design and speaker C's arguing is in opposition (non-reinforcement) to this. However, there are no clear target relations in this passage – it is difficult to consider non-existence to be an alternative to the personal video recorder.

In the future, in addition to the current opinion relations, we would like to define the opinion relations independent of target relations. Consequently, reinforcing/non-reinforcing relations can be made to exist between opinions that do not share target relations or do not even have definite targets.

In this work, we have two types of target relations: the *same* and the *alternative*. These distinctions are a useful first step in capturing the variety of ways in which people express their opinions.

However, all target relationships in the discourse do not necessarily fall into these two categories. It is possible for two items to be related but not have a strictly *same* or *alternative* relation. Thus, in the future we would like to have a general relation category, to capture these cases.

Additionally, creating sub-categories within the *same* relation category will be useful. Let us consider the following example for motivation.

(3.10) C :: **Of course** it **should have** a *on off button*.

B :: Yes , well it **should have** the the the the *expected functionality of a remote control*.

In Example 3.10, the targets *on off button* and *expected functionality of a remote control* are in a *same* target relation as the latter target is a generalization of the on-off feature (the on-off button is a expected functionality). Suppose now, later on in the discourse, the participants speak about the power button. The *power button* is also an expected functionality of the remote, and thus would be in a *same* relation with the target *expected functionality of a remote control*. This puts the targets *on off button*, *expected functionality of a remote control* and the *power button* in the same set. However, the on off button should not be considered to be the same as the power button. This confusion can be avoided

by making the relationships directed and specifying the type of relationship (is-a in this example). Adding such details would prevent inferences such as considering two different types of button as the same.

4.0 ANNOTATION

Our discourse annotation is a stepwise process where text spans revealing opinions and their targets are selected, opinion attributes such as type and polarity are annotated, and finally, targets are linked via one of the two possible target relations. Once these components are annotated, opinion frames exist by definition. That is, in the current scheme, we do not explicitly create annotations for the discourse-level opinion relations.

The full annotation process is carried out in the following stages:

1. Opinion span retrieval: In this step text spans that reveal opinions are annotated.
2. Target span retrieval: In this step text spans that reveal what each opinion is about are annotated.
3. Opinion type annotation: The type attributes (whether an opinion is sentiment or arguing) are then annotated.
4. Opinion polarity annotation: Next, polarity, another opinion attribute (is an opinion positive, negative, neutral or both) is annotated.
5. Target linking: In this stage, links are created between targets that are related in the discourse.
6. Target link labeling: In this last stage, the links created in the previous step are labeled as "same" or "alternative".

We have three goals in this chapter:

Adapting the MPQA subjectivity annotation scheme to meeting data: The MPQA attitude annotation scheme provides guidelines for annotating sentiment and arguing opinions, their polarities and their targets. We conjectured that this scheme would be applicable to our data. Thus, in the case of opinion annotations, our task is that of adapting

the MPQA annotation definitions and scheme for the meeting data. Preliminary experiments revealed that some changes are necessary to the original definitions of the arguing category, to account for peculiarities of the new genre. Once we have adapted and modified the definitions and the annotation policies, we can annotate the data and perform reliability studies.

Creating extensions to the scheme for targets and target relations: After we adapt the MPQA attitude annotation scheme to meetings, we add the discourse-level annotations as extensions. The MPQA scheme provides a definition for targets of opinions. However, we estimated that the target annotations would be different in our case, as we need to capture the target relations. To complete the discourse-level annotations, we design an annotation scheme for annotating the targets and their relations. In order to facilitate the creation of links of different types, we also develop a user-friendly GUI.

Annotating data and performing reliability studies at each stage: Supervised experiments and analysis require reliably annotated data. We also need to test if humans can recognize the elements of the linguistic scheme reliably. As our annotations are applied in a layered fashion, it is necessary to ascertain that the annotations produced at each stage are reliable for the next stage. Thus, we carry out reliability studies at each stage of our annotation process.

Collectively, the different reliability studies in this chapter are aimed to test the high level hypothesis that *the elements of the discourse-level relations defined in this thesis can be reliably annotated by trained human annotators.* This high-level hypothesis covers a number of specific hypotheses, one corresponding to each stage of our annotation described above. These are as follows:

4.a The performance of opinion span retrieval is comparable to that in previous work.

4.b The performance of target span retrieval is comparable to that in previous work.

4.c Human annotators can reliably label types for the given opinion spans.

4.d Human annotators can reliably label polarity for the given opinion spans.

4.e The performance of target linking is comparable to similar previous work.

4.f Human annotators can reliably label the given target links.

For the labeling tasks, we use the Cohen's Kappa metric (κ) [Cohen, 1960] and interpret reliability using the ranges defined by Landis and Koch [Landis and Koch, 1977] and Krippendorf [Krippendorff, 2004]. Note that, for span retrieval tasks we use an information retrieval metric, and hence we compare our results to similar previous work. For the target linking task, we use Passonneau's alpha [Passonneau, 2004], and compare our numbers to that obtained by her for coreference chains.

The rest of this chapter is organized as follows. We describe the adaptation of the concepts of sentiment and arguing from the MPQA annotation scheme and the corresponding reliability experiments in Section 4.1. In Section 4.2, we describe the annotation scheme extensions required for the target relations and the corresponding agreement studies. We discuss some of the challenges faced during annotation in Section 4.3. We conclude with the related work discussion in Section 4.4 and summary in Section 4.5.

4.1 OPINION ANNOTATIONS

In preliminary studies on a separate set of meetings, we found that the definition of sentiment provided in the MPQA annotation scheme is applicable to the meeting data. However, arguing phenomena in face to face conversations was found to be more complex, necessitating a modification in the original definition. Primarily, we found that people can argue for something in a number of ways. The annotation instructions for arguing was extended to explicitly include the wide variety of arguing phenomena in our corpora. Our final annotation scheme explicitly lists the different phenomena, and provides detailed examples to the annotators.

Thus, our formal definitions for the opinion categories are as follows:

- **Sentiment:** Sentiments include emotions, evaluations, judgments, feelings and stances.
- **Arguing:** Arguing includes arguing for/against something, arguing that something is true/false, or should/should not be done. It brings out the participant's strong conviction and/or his attempt to convince others. Arguing also includes expressing support for (or against) or backing the acceptance of an object, viewpoint, idea or stance by providing

reasoning, justifications, judgment, evaluations or beliefs. This support or backing may be explicit or implicit[1].

As a result of the extensions, our arguing annotations not only capture explicit assertions and expressions of convictions, but it also captures arguing expressed as necessities, persuasive constructs, justifications or communal desires. Examples 4.1 to 4.7 illustrate the various ways in which people argue in our face to face conversational data (which the new arguing definition captures).

(4.1) **I think** this idea will work

(4.2) This is the lightest remote **in the world**

(4.3) We **ought** to get this button

(4.4) **Clearly**, we cannot afford to use speech recognition

(4.5) It would be nice **if we could** have the curved shape

(4.6) I brought this up **because** this will affect the cost

(4.7) **We want** a fancy look and feel

In Example 4.1, the speaker argues by explicitly stating his conviction. In Example 4.2, the speaker simply asserts his argument using the exaggeration. The exaggeration is used to draw emphasis to his viewpoint. In Example 4.3 the speaker argues for getting the button by framing it as a necessity. In Example 4.4, the speaker states his proposition categorically to argue for it. Interestingly, in face to face conversations, participants also use persuasive constructs, justifications or communal desires to argue for something as in Examples 4.5, 4.6 and 4.7, respectively. We found that context, in addition to lexical clues is needed to infer that arguing is taking place. As part of a casual conversation, the utterance "I think John was at home" would not be arguing, despite the presence of the lexical anchor "I think". However, in a debate about John's whereabouts at the time of a murder, the sentence could function as arguing. Here the context and the knowledge that there is a disparity between the speakers helps us infer that the sentence is intended to argue.

[1]In this work we handle only explicit arguing.

	Sentiment	Arguing
segments	0.826	0.716
sentences	0.789	0.677

Table 2: Kappa values for Inter-annotator agreement in detecting Opinions

4.1.1 Preliminary Reliability Study At The Segment And Sentence Level

After modifying the annotation definitions, we performed a reliability study to test the whether our opinion types can be detected in the meeting data. This study is performed at the coarse granularity of the utterance (segment) and the sentence[2], and tests whether performing opinion analysis in the meeting genre is promising at all.

Two annotators (one of them is the author) underwent 3 rounds of training. Then we calculated inter-annotator agreement using Cohen's Kappa [Cohen, 1960] over a previously unseen meeting (607 segments, 1002 sentences). Although the annotators tag expressions, in this study, agreement is calculated over the segment or the sentence. For this purpose, we assign a segment (or sentence) the labels of all the expressions annotated within it. If a sentence (or segment) has sentiment as well as arguing annotations, it is assigned both these labels. This does not affect our evaluations as we evaluate each category separately.

Table 2 shows the results of the agreement study. Our inter-annotator kappa values are in the substantial agreement range according to Landis and Koch [Landis and Koch, 1977], which indicates that the annotations are reliable at the segment and sentence level. According to Krippendorf's interpretation, our kappa values are in the tentatively reliable range (the reliability of sentiment annotation at the segment level is in the good reliability range). Compared to sentiment, arguing has lower kappa values at 0.716 at the segment and 0.677 at the sentence level. This is in line with our intuition that the arguing category is inherently difficult.

[2]We use the automatically detected segment and manually annotated sentence information provided in the AMI corpus.

4.1.2 Final Opinion Annotation Scheme

The final opinion annotation is composed of the following components:

- **Opinion Span** This is the span of text that reveals the opinion. This may be a single word or multiple-word phrases.
- **Type** This is an attribute that specifies if the opinion is a *sentiment* or *arguing.*
- **Polarity** The polarity attribute captures the valence of the opinion expression in context; and it can have 4 possible values: *positive*, *negative*, *neutral*, *both.*
- **Target(s)** The target attribute is used to capture what the opinion is about. An opinion has zero or more targets.

We should note here a difference in the opinion span annotation policy between our scheme and the previous scheme [Wilson, 2008b, Wilson and Wiebe, 2005]. Our policy for opinion span annotation instructs the annotators to mark the minimum text span that conveys the opinion. Specifically, we focus on the lexical anchors that convey sentiment and arguing opinions. Using this approach, the annotations for sentiment text spans for meetings are similar to that obtained by Wilson for news. However, our approach of marking the lexical triggers for arguing resulted in significantly different spans of annotations for the arguing category. To give an example on how the annotation of spans would differ, let us consider this example from Wilson's dissertation [Wilson, 2008b]

(4.8) Iran insists its nuclear program is purely for peaceful purposes.

The annotation span for arguing in Wilson's scheme is "insists its nuclear program is purely for peaceful purposes", whereas in our scheme the annotation span is "insists". Our annotation would mark "its nuclear program is purely for peaceful purposes" as a target of the arguing.

Another difference in our annotation policy is that the polarity of an opinion is always annotated with respect to its target (if any). This requires that the target of an opinion be determined before its polarity is annotated. We introduced this policy because, in our preliminary experiments, we found that depending on what the target span (and the opinion span) is perceived to be, the polarity can change. Let us consider the two possible interpretations of the sentence "We need to get rid of the LCD display." to illustrate this.

(4.9) We need to get rid of the LCD display

(interpretation-1) **We need to** *get rid of the LCD display.*

(interpretation-2) **We need to get rid of** the *LCD display.*

According to the first interpretation, the opinion is **We need to** and its target is *get rid of the LCD display.* In this case, the annotation conveys a positive opinion towards the idea of getting rid of the LCD display. In the second interpretation, the whole span **We need to get rid of** is a negative arguing, and this negative arguing is directed towards the object, the *LCD display.* Notice that, in both interpretations, the annotations capture the same underlying semantics, but the polarity is the opposite. In order to circumvent such errors, we devised the policy that the polarity of an opinion is determined with respect to its target. Thus, we perform target span annotations prior to polarity annotations (in stage 2). However, we do not provide any details for the targets (their relations, etc.).

4.1.3 Annotation Tool

In order to produce the annotations, we use the GATE annotation tool [Cunningham et al., 2002]. GATE has been employed successfully in previous work on subjectivity and attitude annotations. GATE provides a user interface for producing text span annotations. Using the interface, annotators can select specific spans of text that reveal opinions or targets. Each of the opinion span annotations has type, polarity and target (id) attributes that are populated. The annotation types and attributes to be displayed are supplied to GATE using an XML schema. GATE can store annotations as byte offset annotations as well as XML. We use byte offset format to be compatible with the MPQA data.

4.1.4 Agreement Studies

In this section, we measure the reliability of marking opinion spans and annotating the type and polarity attributes. The reliability study is reported as follows: opinion expression annotation (Section 4.1.4.1), type annotation (Section 4.1.4.2) and polarity annotation (Section

Gold	Exact	Lenient	Subset
ANN-1	53	89	87
ANN-2	44	76	74

Table 3: Inter-Annotator agreement on Opinion Span retrieval

4.1.4.3). We have two annotators for this study and one of them is the author. The annotators underwent three rounds of training for each stage. This agreement study for each stage was performed over about 250 utterances.

4.1.4.1 Reliability study for opinion span retrieval This inter-annotator agreement study tests the hypothesis that the performance of opinion span retrieval is comparable to that in previous work. In this study, the annotators selected text spans and labeled them as *opinion*. We do not make the type distinction here, as we wanted to study the opinion span retrieval in isolation.

We calculate our agreement for text span retrieval similar to Wiebe et al. [Wiebe et al., 2005]. This agreement metric corresponds to the precision metric in information retrieval, where annotations from one annotator are considered the gold standard, and the other annotator's annotations are evaluated against it.

Table 3 shows the inter-annotator agreement (in percentages). For the first row, the annotations produced by annotator-1 (ANN-1) are taken as the gold standard and for the second row, the annotations from annotator-2 form the gold standard. In particular, the precision for the first row is calculated as:

$$precision(ANN2||ANN1) = \frac{|ANN1\ matching\ ANN2|}{|ANN2|}$$

and the precision for the second row is calculated as:

$$precision(ANN1||ANN2) = \frac{|ANN1\ matching\ ANN2|}{|ANN1|}$$

Note that we could have instead used the recall metric, in which case the two formulas and the numbers in the two rows would have been interchanged. That is,

$$recall(ANN2||ANN1) = precision(ANN1||ANN2)$$

The "Exact" column in Table 3 reports the agreement when two text spans have to match exactly to be considered correct. The "Lenient" column shows the agreement under the condition that an overlap relation between the two annotators' retrieved spans is also considered to be a hit. Wiebe et al. [Wiebe et al., 2005] use this approach to measure agreement for a (somewhat) similar task of subjectivity span retrieval in the news corpus. Our agreement numbers for this column are comparable to theirs. This result indicates that the reliability of our opinion span annotation in the meeting genre is comparable to subjectivity span annotation in monologic texts.

Finally, the third column, "Subset", shows the agreement for a more strict matching. Here, one of the spans must be a sub-span of the other to be considered a match. Two opinion spans that satisfy this condition are ensured to share all the opinion words of the smaller span. The relatively higher agreement numbers for this column, as compared to the Exact column, indicate that while the annotators do not often retrieve the exact same span, they reliably retrieve approximate spans. Interestingly, the agreement numbers between Lenient and Subset columns are close. This implies that, in the cases of inexact matches, the spans retrieved by the two annotators are still close – they agree on the opinion words and differ mostly on the inclusion of function words (e.g. articles) and observation of syntactic boundaries.

4.1.4.2 Reliability study for opinion type annotation

The experiments in this section test whether annotators can label opinion categories reliably in the meeting data. That is, whether they are able to distinguish between the two opinion categories.

In order to isolate errors at this stage from errors occurred in the previous stage (secondary errors), we first perform consensus annotation to detect the opinion spans. These opinion spans are provided to the annotators, who then annotate these spans for the opinion

	Type Tagging
Accuracy	97.8%
κ	0.95

Table 4: Inter-Annotator agreement on Opinion Type annotation

	Polarity Tagging
Accuracy	98.5%
κ	0.952

Table 5: Inter-Annotator agreement on Opinion Polarity annotation

type. As every opinion instance is tagged with a type, we use accuracy and Cohen's Kappa (κ) metric [Cohen, 1960]. Accuracy is calculated as

$$Accuracy = \frac{|mismatch|}{|instances|}$$

The results, reported in Table 4, show that κ both for opinion type annotations is high. Specifically, using Krippendorf's interpretation of kappa, the reliability falls in the "good reliability" range. Using the interpretation of kappa by Landis and Koch, our reliability falls in the "almost perfect agreement" range. This indicates that sentiment and arguing can be reliably distinguished once the opinion spans are known.

4.1.4.3 Reliability study for opinion polarity annotation

The experiments in this section test whether opinion polarity can be reliably detected in the meeting data.

In order to isolate errors incurred at this stage from secondary errors, we provide the annotators with opinion span annotations. As opinion polarities are judged with respect to their targets, the annotators are also provided with target annotations. As every opinion instance is tagged with polarity, we use accuracy and Cohen's Kappa (κ) metric [Cohen, 1960] to measure agreement. The results, reported in Table 5, show that κ for polarity tagging

is high. Specifically, using Krippendorf's interpretation, the reliability falls in the "good reliability" range and using the interpretation by Landis and Koch, our reliability falls in the "almost perfect agreement" range. Our polarity detection task shows an improvement in κ over a similar polarity assignment task by Wilson et al. [Wilson et al., 2005a] for the news corpus (κ of 0.72). We believe this improvement can partly be attributed to the target information available to our annotators.

4.2 TARGET AND TARGET RELATION ANNOTATION

In this section we discuss the annotation of the target spans and the creation of the target links and their labels. The target annotation consists of the following attributes:

- **Target Span:** This is a span of text that captures what an opinion is about. This can be a proposition or an entity.
- **Target ID:** Every new target annotation is provided with a unique ID. This ID is used to link the targets to their respective opinions and also link targets to each other. The ID assignment is automatically done by our GUI annotation interface for every new target span.
- **Target Link:** This is an attribute of a target and records all the targets in the discourse that the target is related (linked) to.
- **Link Type:** The relation type between two targets is specified by this attribute and its value is either *same* or *alternative*. For every link listed in the **Target Link** attribute above, there is a corresponding entry in this attribute field.

The annotation process is carried out in the same order as the attributes are listed above. Once the target span is annotated and assigned an ID, the target is tied to its opinion. This is done by entering the target ID in the "target" attribute field of the opinion annotation. In the succeeding annotation stage, if two targets are related, this information is annotated by populating the **Target Link** field of one of the targets with the ID of the other target. The **Link Type** entry is then made corresponding to each link.

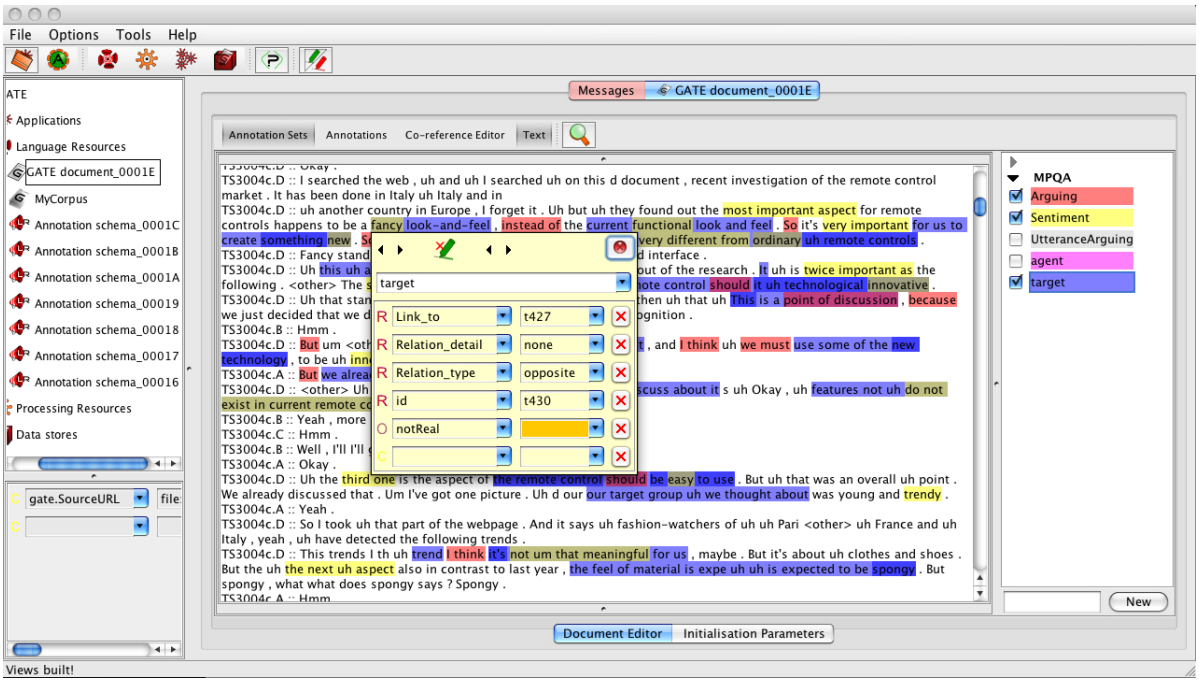

Figure 6: Annotation Tool - GATE (regular) view

Links between targets can be followed in either direction to construct chains. In this work, we consider target relations to be commutative (undirected), i.e., Link(t1,t2) => Link(t2,t1). When a newly annotated target is similar (or opposed) to a set of targets already participating in *same* relations, then the *same* (or *alt*) link is made only to one of them – the one that looks most natural. This is often the one that is closest.

4.2.1 Annotation Tool

The GATE annotation tool [Cunningham et al., 2002] is effective for creating individual opinion annotations. However, GATE does not provide a way to create and label pair-wise relations between text-span annotations.[3] In our work, we need to overlay the link annotations over the span-based target annotations.

Additionally, when a new target is annotated, the annotator has to check if it is *same*

[3]GATE does have facilities to create coreference chains, but our task requires *different types* of relations.

or *alternative* to any of the previously annotated targets. If a new target is *same* as a set of targets already related via *same* target relations, the annotator has to connect it to at least one of the targets in that set. Thus, it is necessary for annotators to be able to easily inspect and navigate the existing target chains. As AMI meetings generally have over 1000 sentences, the cognitive load on annotators increases tremendously as the annotation progresses, especially because long-distance relations are possible. Thus, we created a GATE plugin, called the *link-GUI*, to provide a link annotation environment in GATE. This tool also provides annotators with a visual representation of the links associated with a given target, and easy accessibility to all the related targets in the document.

Functionally the link-GUI is loaded along with the regular GATE interface. They provide different views of a given document and its annotations. The regular GATE interface is shown in Figure 6, and the link-GUI interface that provides the discourse-relation view for the same document is shown in Figure 7. Notice that link-GUI displays the relations for one target at a time. The annotations produced in either of the views is synchronized real-time, and can be inspected in the other view. Thus, the annotator can switch back and forth between the two views seamlessly.

The annotator performs opinion and target span annotations using the regular GATE interface. That is, opinions and their attributes (including type and polarity), as well as targets are annotated via the regular interface. When she is ready to create links between the targets, she uses the link-GUI interface. Figure 6 shows details of the target annotation for the target "fancy look and feel" in the regular GATE view. The attribute pop-up window indicates that this target is linked to another target that has an ID t427, and the relation type is alternative[4]. Note that these attributes were populated using the link-GUI interface.

Figure 7 shows the view in the link-GUI for the same target ("fancy look and feel"). The panel on the bottom left reproduces the meeting snippet with details of the discourse relation highlighted. The target being inspected, "fancy look and feel", is highlighted in purple, its opinion ("most important aspect") is highlighted in blue, the related target ("current functional look and feel") is in orange and its opinion ("functional") is in red. The panel on the bottom right shows the target chain in which the given target participates. All targets

[4]For historical reasons, the alternative is denoted as "opposite" in the schema.

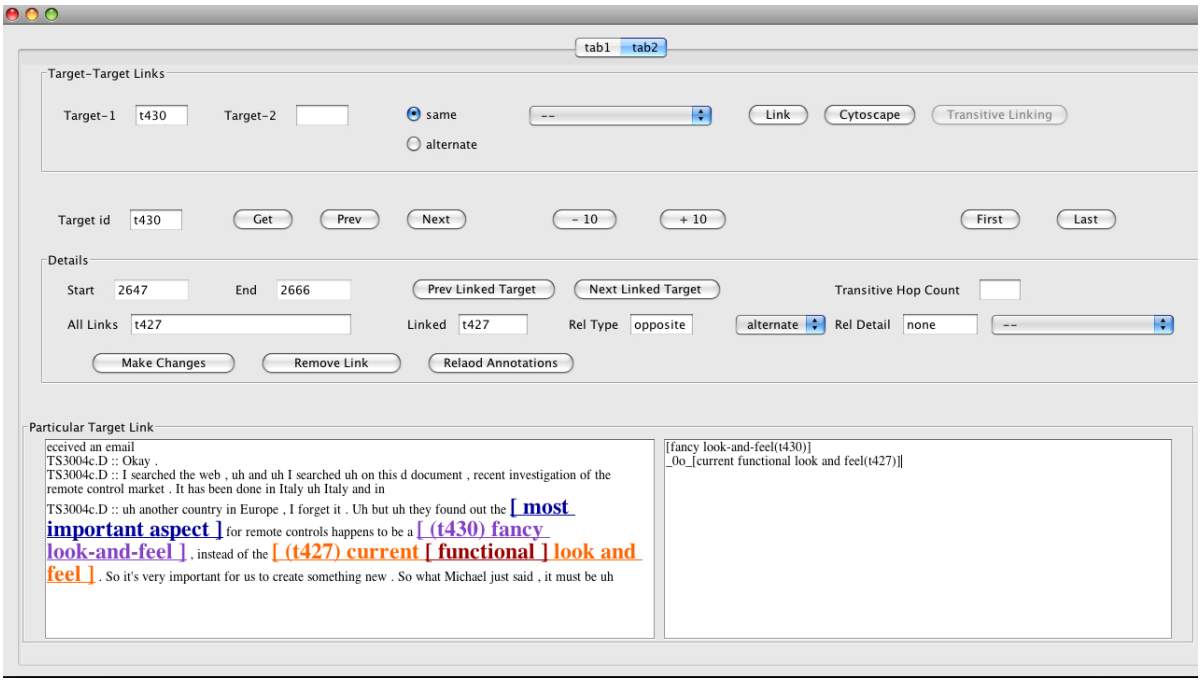

Figure 7: Annotation Tool - relation view

that can be reached from this target, either directly by annotated links or by following links via other intermediate targets, are listed in this window. As we can see in Figure 7, this window shows that the target Target-1 ("fancy look and feel") is related to target "current functional look and feel" via a direct alternative link. The "0" on the link denotes that it is a direct link (hop count of zero), and the "o" indicates that the relation is of type alternative (opposite).

The top and the middle panels of link-GUI interface are used for entering the relation annotation and for navigating over the existing annotations.

4.2.2 Agreement Studies

The agreement studies in this section are presented in three stages. The first stage is the annotation of the target span, the second is the target relation detection (populating the Target Link attribute for each target) and the third stage is the labeling of the relation

Gold	Exact	Lenient	Subset
ANN-1	74	87	87
ANN-2	76	90	90

Table 6: Inter-Annotator agreement on Targets with Perfect Opinion spans

type for each link from the second stage as *same* or *alternative* (populating the Link Type attribute of the target annotation). In order to prevent errors incurred at earlier stages from affecting the evaluation of later stages, the annotators produced a consensus version at the end of each stage, and used this consensus annotation as the starting point for the next annotation stage. We have two annotators (one of them is the author), and three rounds of training for each stage. The reliability tests are carried out over (about) 250 utterances (half a meeting).

4.2.2.1 Reliability study for target span annotation

The experiments in this section test whether target spans can be reliably annotated.

Targets are annotated for all opinions in the corpus. Thus, opinion span annotations are performed prior to target span annotations. Depending on the span chosen for opinion annotation, the target span annotations in the following stage can vary. In Example 4.9 we observed that the choice of the opinion span changed the span of the target. Hence, in order to isolate the secondary errors, that is, errors form the opinion annotation stage, we perform evaluations for targets of the opinions on which the annotators perfectly agree. Thus, given a corpus with opinion span annotations, the annotator's task in this stage is to select the spans that reveal the targets.

The evaluation for target span annotation is similar to opinion span retrieval – the agreement metric used corresponds to the precision metric in information retrieval, where annotations from one annotator are considered the gold standard, and the other annotator's annotations are evaluated against it. Table 6 shows the inter-annotator agreement (in percentages). For the first row, the annotations produced by Annotator-1 (ANN-1) are taken

as the gold standard and, for the second row, the annotations from annotator-2 form the gold standard. The precision for the first row is calculated as:

$$precision(ANN2||ANN1) = \frac{|ANN1\ matching\ ANN2|}{|ANN2|}$$

and the precision for the second row is calculated as:

$$precision(ANN1||ANN2) = \frac{|ANN1\ matching\ ANN2|}{|ANN1|}$$

Some researchers also use the recall metric where the recall for the first row is calculated as:

$$recall(ANN2||ANN1) = \frac{|ANN1\ matching\ ANN2|}{|ANN1|}$$

and the recall for the second row is calculated as:

$$recall(ANN1||ANN2) = \frac{|ANN1\ matching\ ANN2|}{|ANN2|}$$

Notice that the numbers for precision and recall indicate the same thing –

$$precision(ANN2||ANN1) = recall(ANN1||ANN2)$$

and

$$precision(ANN1||ANN2) = recall(ANN2||ANN1)$$

Thus, we report only the precision metrics here. The "Exact" column in Table 6 reports the agreement when two text spans have to match exactly to be considered correct. The "Lenient" column shows the results if an overlap between the two annotators' retrieved spans is also considered to be a hit. Finally, the third column, "Subset", shows the agreement for a more strict constraint: one of the spans must be a sub-span of the other to be considered a match. This target span retrieval study is similar to that conducted by Wilson [Wilson, 2007] for target frames associated with attitudes. Wilson uses the recall metric. As noted above, the precision recall metrics in this evaluation can be used interchangeably. The agreement numbers obtained by us in Table 6 (87%-90%) are comparable to the agreement obtained by Wilson (86% - 87%) on monologic texts.

Meeting:	a	b	c	d
Target linking (α)	0.79	0.74	0.59	0.52

Table 7: Inter-Annotator agreement on Target relation identification

4.2.2.2 Reliability study for target linking The experiments in this section test whether target relations can be reliably detected. For this, a meeting annotated with elements of all the previous stages are provided to the annotators. The annotators determine which of the given targets should be related and create (unlabeled) links.

For evaluating target linking, we treat target links in the discourse similarly to coreference chains and apply evaluation methods developed for coreference resolution [Passonneau, 2004]. Passonneau's method is based on Krippendorf's α metric [Krippendorff, 2004] and allows for partial matches between anaphoric chains. Passonneau [Passonneau, 2004] reports that, in her coreference task on spoken monologs, α varies with the difficulty of the corpus (from 0.46 to 0.74). Though our tasks are not directly comparable, our results too show a similar trend. Table 7 shows our agreement for the four types of meetings in the AMI corpus: a kickoff meeting (a), a meeting discussing functional designs of the TV remote (b), a meeting discussing conceptual designs (c) and a meeting discussing detailed design and final prototype evaluation (d). Of the meetings, the type (a) data we use has relatively clear discussions. Our type (c) meeting is the most difficult, as participants are expressing opinions about hypothetical (desirable) TV remotes. In our detailed design meeting (d), there are two final designs being evaluated. On analyzing the chains from the two annotators, we discovered that one annotator had maintained two separate chains for the two remotes as there is no explicit linguistic indication (within the 250 utterances) that these two are alternatives. The second annotator, on the other hand, used the knowledge that the goal of the meeting is to design a single TV remote to link them as alternatives. Thus, by changing just two links in the second annotator's file to account for this, our α for this meeting went up from 0.52 to 0.70.

Meeting:	a	b	c	d
Relation Labeling (κ)	1	1	0.91	1

Table 8: Inter-Annotator agreement on Target link labeling

4.2.2.3 Reliability study for link labeling The experiments in this section test whether the target relations detected in the previous stage can be labelled reliably. This study tests whether annotators are able to distinguish between the *alternative* and *same* relations. For this stage, a corpus with unlabeled target relations is provided to the annotators which eliminates secondary errors. We evaluate type (*same*/*alternative*) labeling task with the help of the κ metric. Table 8 reports the κ values.

The high κ for the relation type identification shows that once the presence of a link is detected, it is not difficult to determine if the targets are similar or alternatives to each other. The reliability of the link labeling task is in the "good reliability" range using Krippendorf's interpretation, and in the "almost perfect agreement" range using Landis and Koch's interpretation. This indicates that the difficulty in the target linking lies in determining if two targets are related in the discourse. Once that decision is made, labeling of the links is an easy task.

4.3 DISCUSSION

Our agreement studies help to identify the aspects of our annotation that are straightforward, and those that need complex reasoning. In this section, we will discuss some of the challenges faced during the annotation process.

Our results indicate that while opinion type, opinion polarity and target relation type can be reliably produced by humans, retrieval of opinion spans, target spans and target links is relatively more difficult. Specifically, our results show that annotators do not usually agree on the exact spans, but are in relatively more agreement on the approximate spans. This

is usually due to the inclusion of articles and observation of syntactic boundaries by one annotator and not the other. Additionally there are also situations where there is ambiguity on what is the minimal text span that reveals the opinion. This situation is more prevalent with the arguing category. Example 4.10 illustrates this. One annotator considers the text span "have to" sufficient to reveal the arguing, while the other considers the whole clause "have to be taken into account" necessary to capture the arguing.

(4.10) D :: Yeah *these numbers* **<have to> be taken into account**

A common cause of annotation disagreement is different interpretations of an utterance, particularly in the presence of disfluencies and restarts. For example, consider the following utterance where a participant is evaluating the drawing of another participant on the white board.

(4.11) *It's a baby shark* , **it looks to me**, ...

One annotator interpreted "it looks to me" as an arguing for the belief that it was indeed a drawing of a baby shark (positive arguing). The second annotator on the other hand looked at it as a *neutral* viewpoint/evaluation (sentiment) being expressed regarding the drawing. Thus, even though both annotators felt that an opinion is being expressed, they differed on its type and polarity.

There are some opinions that are inherently on the borderline of sentiment and arguing. For example, consider the following utterance where there is an appeal to importance:

(4.12) **Also important** for you all is um the *production cost must be maximal twelve Euro and fifty cents.*

Here, "also important" might be taken as an assessment (sentiment) of the high value of adhering to the budget (relative to other constraints), or simply as an argument for adhering to the budget.

One potential source of problems to the target-linking process consists of cases where the same item becomes involved in more than one opposition. For instance, in the example below, speaker D initially sets up an alternative between speech recognition and buttons as a possible interface for navigation. But later, speaker A re-frames the choice as between having

speech recognition only and having both options. Connecting up all references to speech recognition as a target respects the co-reference but it also results in incorrect conclusions: the speech recognition is an alternative to having both speech recognition and buttons.

(4.13) A:: One thing is **interesting** is talking about *speech recognition* in a remote control...

D:: ... So that we don't need any button on the remote control it would be all based on speech.

A:: ... I think *that* **would not work so well**. **You wanna** have *both options*.

The nature of meetings creates an additional ambiguity which causes annotation errors. In face to face meetings, participants establish common ground by repeating the previous speakers words. Participants also have to summarize and document the proceedings. Hence, they repeat other members' utterances, which include previously expressed opinions. It is not always clear if this repetition is due to alignment between the speakers' opinions or for purely documentation purposes. Example 4.14 illustrates this situation. Here, it is not clear if the "too much time" in speaker D's utterance reflects his opinion, or if it is repeated during documentation. Depending on the annotators' judgment of the situation, this span will be annotated as an opinion.

(4.14) A :: Thirty four percent said it took **too much** *time to learn to use a new one.* Yep.

D :: Okay too much time to learn. Okay.

The annotation difficulty is also increased due to the multi-modal nature of the meetings. Annotators listened to speech signals for the meetings while annotating, which resolved some ambiguities. In fact, we found that listening to the speech in addition to reading the transcription improves the annotation reliability [Somasundaran et al., 2006] of subjectivity categories. However, in addition to text and speech, gestures and visual diagrams play an important role in face-to-face AMI meetings. In the absence of visual input, annotators would need to guess what was happening.

4.4 RELATED WORK

The work most close to our opinion annotation work is the subjectivity annotation by Wilson [Wilson, 2008a] on the AMI meeting corpus. Our work predates this work. Also, the categories of opinions and the definition of targets for Wilson's annotation scheme is different from ours. The scheme in [Wilson, 2008a] is designed with a particular application in mind – the meeting assistant. The high-level annotation categories are Subjective and Polar Objective. Targets belong to a pre-specified list and denote what specific item of the meeting the subjective (or polar objective) utterance is directed toward; for example, is the utterance directed toward the meeting or the remote design.

Previous researchers have generated corpora with annotations of positive and negative sentiments at the sentence level [Yu and Hatzivassiloglou, 2003, Bethard et al., 2004, Kim and Hovy, 2004]. Our opinion annotations are at the expression level. Additionally, we annotate arguing opinions. Furthermore, our work is on face to face meeting data. Hu and Liu [Hu and Liu, 2004b, Hu and Liu, 2004a] annotate products and product features, which are similar to our target annotations. However, our target annotations are more general – while our targets may be products or product features, they may also be propositions, events or anything that the opinion is about.

In this work, we build on the MPQA annotation scheme, which is based on subjectivity theory. Appraisal theory [Martin and White, 2005] analyzes text based on affect (emotional response), judgment (evaluations) and appreciation (used for evaluating products and processes) and is a theory that runs parallel to subjectivity theory. Appraisal theory has some overlap with the core subjectivity theory and its attitude extensions. Even though there is some overlap, these two schemes are inherently different. As explained in [Wilson, 2007], the subjectivity attitude scheme does not distinguish between affect and evaluations; these fall under the sentiment category.

This work is the first attempt in annotating subjectivity categories in multi-party conversations. Previous work in speech-based data has attempted to detect emotions such as frustrations, annoyance, anger, happiness, sadness, and boredom (for e.g., [Liscombe et al., 2003, Devillers et al., 2005, Litman and Forbes-Riley, 2006]). Our work is not concerned with the

speaker's emotions, but rather opinions toward the issues and topics addressed in the meeting.

"Hot spots" in meetings [Wrede and Shriberg, 2003] relate to our work because they find sections in the meeting where participants are involved in debates or high arousal activity. While that work distinguishes between high arousal and low arousal, it does not distinguish between opinion or non-opinion or the different types of opinion.

There are efforts to annotate the mental states of participants in meetings or interviews on the basis of multi-modal data [Devillers et al., 2005, Reidsma et al., 2006]. The focus of these kinds of research is different from ours in that they target the actual mental states of the speakers in the unfolding situation, while we focus on subjective states communicated through language. While often the same, they are not necessarily identical as language allows for displacement: participants may calmly report about other people's anger, report their past or expected future mental states, etc.

Researchers are also annotating the decisions in the meetings [Hsueh and Moore, 2007, Purver et al., 2006]. While our annotations track opinions in the decision making process, the decision detection research is mostly concerned with its outcome. Dialogs have also been annotated with discourse schemes such as RST [Stent, 2000]. As mentioned in Chapter 3, our discourse-level relations are different from RST-based relations.

4.5 SUMMARY AND FUTURE WORK

In this chapter, we described the annotation of AMI meeting data with our discourse-level relations. In this process, we started from the MPQA attitude annotation scheme, adapted it to the meeting genre, and extended it to capture our discourse-level relations.

Our annotation effort is the first in subjectivity-based annotations in the meeting genre. We found that the difference in genres does not affect the sentiment category, but the arguing phenomena is more varied. To account for the wide variety of ways in which arguing is performed, we extended the arguing definition. We defined extensions to the annotation scheme to capture target relations. We also developed a GUI-based annotation plugin for

GATE in order to facilitate discourse-level linking.

We carried out a number of reliability studies. Our opinion span retrieval reliability is comparable to the reliability of subjectivity span retrieval in monologic texts [Wiebe et al., 2005], which supports our hypothesis 4.a. Our annotation reliability for polarity tagging, opinion type tagging and target link labeling tasks are in Krippendorf's "good reliability" and Landis and Koch's "almost perfect agreement" ranges. These results support hypotheses 4.c, 4.d and 4.e. Similarly, our target span annotation reliability is comparable to the reliability of target annotations in monologic texts, thus supporting hypothesis 4.b. Finally, for target linking, our alphas are in the similar range as that reported by Passonneau, which provides support for hypothesis 4.e. Overall, our results indicate that the elements of the discourse-level relations defined in this thesis can be identified reliably by trained annotators.

Our experiments provide a number of insights: we discovered that annotators tend to agree on opinion words, but depending on the observance of syntactic boundaries, the spans may differ. We found that target linking is a difficult task and its reliability varies depending on the difficulty of the data. In general we found that, in the process of creating the discourse-level relations, finding opinion and target spans and detecting target relations are relatively challenging, while labeling tasks such as polarity tagging, type tagging and link labeling are relatively unambiguous.

We presented a discussion of the different challenges faced by annotators in this data. We observed that disfluencies and restarts, as well as the nature of the meetings, adds complexities to the annotation process.

The annotation scheme developed in this chapter is aimed at capturing the discourse-level relations specified in the linguistic scheme. Thus, it focuses on creating the elements that comprise the opinion frames. In future, we would like to extend this annotation scheme to capture finer distinctions in the target relations corresponding to the extensions to the linguistic scheme proposed in Chapter 3.

The current annotation scheme is flat, in that all opinions regarding essentially the same concept are encoded using the *same* relations. However, there are situations where a hierarchical annotation structure will be useful. For instance, in the Example 1.2, "that"

refers to the rubbery material being durable, and not directly to the rubbery material. Due to our flat structure, we treat "that" similar to the other anaphoric references in the passage. By making the annotations hierarchical, we will be able to capture finer nuances of discourse-level opinion relations useful for supporting overall stances. Such explorations would be a part of future work.

In the current annotation scheme, annotators read the speech transcription and listened to the accompanying speech. However, this does not cover all the modalities of face-to-face conversations. In the future, we would like to add video to the information available during the annotation process, which promises to further resolve ambiguities.

5.0 FINE-GRAINED POLARITY DISAMBIGUATION

In the previous chapters we presented a scheme and its annotation for our discourse-level relations. This chapter is the next step – employing these discourse-level relations to improve opinion analysis. Specifically, we will focus on fine-grained opinion polarity classification. We endeavor to test the hypothesis that *the discourse-level relations are useful for fine-grained polarity disambiguation.*

In this work, we also investigate *how* the discourse-level information can be incorporated effectively for fine-grained polarity classification. That is, we explore computational modeling choices that will achieve an overall, discourse-coherent inference ("global" inference). This is done by incorporating discourse-based information in addition to information available in the local context. The local context ("local" information) comprises of the words in the instance[1] to be classified. We use a two step approach for global inference: first, a word-based classifier produces polarity classification for each instance locally, and in the second step the global inference is applied. We explore two global inference paradigms: an optimization framework and a collective classification framework. Specifically, we have the following goals:

Exploring linguistic features for local polarity classification: In our two step global inference approach, the first step is polarity classification using local information. The discourse-based methods bootstrap using this classification. Thus, it is important that the local classifier be reliable, otherwise noise will propagate in the global classification step. We explore a number of linguistic features to achieve reliable polarity classifications.

Computationally modeling the discourse relations using an unsupervised optimization framework: The optimization framework encodes, via constraints, human in-

[1] An instance can be a sentence, segment (utterance) or any other text span.

sights about discourse coherence involving our discourse-level relations. This formulation does not require the global system to learn an interdependent polarity classification from the annotations, and is thus unsupervised. However, it does require human effort for creating the constraints.

Computationally modeling the discourse relations using a collective classification framework: The collective classification framework is supervised – an interdependent, global polarity classification has to be learnt by the classifier from the annotated data. This system does not require human insights regarding discourse coherence to be formally encoded into the system. However, it does require annotated data.

Given our investigation of the two different modeling paradigms, we have two very specific hypotheses in addition to the main hypothesis state above. These are:

5.a Discourse-level relations can be effectively modeled using an optimization framework to produce classification improvements over the approach that uses local information alone.

5.b Discourse-level relations can be effectively modeled using a collective classification framework to produce classification improvements over the approach that uses local information alone.

The rest of this chapter is organized as follows. We will first introduce our data for this work in Section 5.1, and explain the general framework for our global inference methods in Section 5.2. Section 5.3 describes the word-based local classifier, Section 5.4 describes the global optimization method and Section 5.5 describes the global collective classification approach. Our experiments with these frameworks are presented in Section 5.6 and discussion in Section 5.7. We talk about related work in Section 5.8 and conclude in Section 5.9.

5.1 DATA

We use Dialog Act (DA) segmentation of the meeting corpus for our experiments. These segmentations are provided with the AMI data. Dialog Act units are text spans that convey

a dialog-level intent. We chose DA segmentation over sentence-based segmentation because preliminary inspection of the data revealed that the DAs are less likely to contain multiple expressions of opinions.

Figure 8 shows the text span annotations for Example 1.2. We see that this passage has two sentences. The first sentence contains one opinion expression, and the second contains three opinion expressions. There are two *same* target relations and two reinforcing opinion frame relations between the first and second sentences. Also, there is a *same* target relation (and a reinforcing opinion frame relation) between text spans *within* the second sentence.

Figure 9 shows the same example with DA segmentation. The two sentences in the original passage are split into 5 DA units (shown in the figure as the red boxes DA-1 to DA-5). Notice here, that the structure is relatively more neat. First, there is a reduction in the number of opinion expressions contained in each unit. As a result, there is also a reduction in the number of links within each unit. Also observe that there are no multiple target relations or opinion frame relations between the units. As we see, the DA segmentation is more convenient and hence we use DAs as units for our polarity classification.

As DAs are our units for classification, we map opinion annotations, which are text-span based, to the DA units. If a DA unit contains an opinion annotation, the polarity label of that opinion is transferred upwards to the containing DA. When a DA contains multiple opinion annotations, each with a different polarity, one of them is randomly chosen as the label. However, our choice of DAs as classification units mitigates the number of cases where such a choice has to be made.

Let us explain the process of label transfer between the text spans and their containing DAs in Figure 9. The polarity annotation "positive" for the text span **a bit more bouncy** is transferred upwards to DA-1. That is, DA-1 has a positive polarity label. Similarly, DA-4 and DA-5 have positive polarity labels. DA-1 and DA-3 do not contain opinions. Hence, they are assigned the default "neutral" label. In this work, DAs containing opinions with neutral polarity, or with no opinion content are considered neutral.

Discourse-level target relations are also transferred upwards, form the text spans to the containing DAs. This is illustrated in Figure 9. DA-2 and DA-4 are in a *same* target relation, as the targets in them are in that relationship. Similarly DA-4 and DA-5 are in a

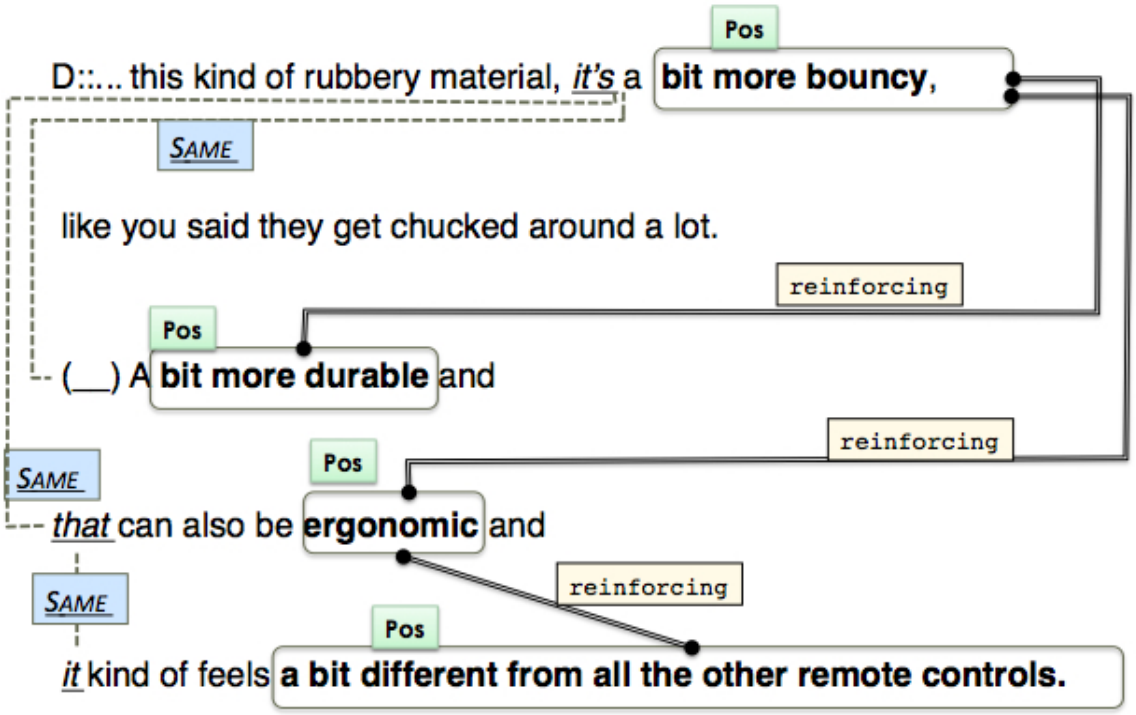

Figure 8: Original Text span annotations for Example 1.2

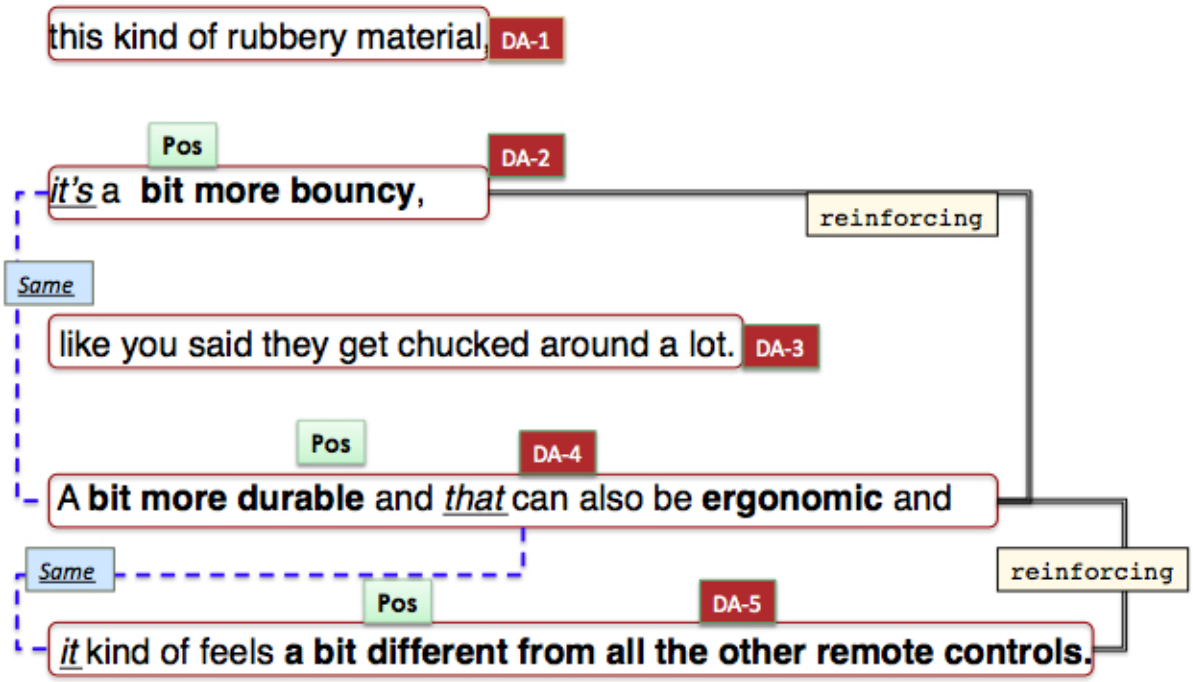

Figure 9: Example 1.2 with Dialog Act segmentation and transferred annotations

same target relation. Recall that, when a newly annotated target is similar (or opposed) to a set of targets already participating in *same* relations, then the annotators are required to make the *same* (or *alternative*) link only to one of them – the one that looks most natural. This is often the one that is closest. This explains why there is no direct relation between DA-2 and DA-5. However, they are indirectly connected via DA-4.

Discourse-level opinion relations are also transferred upwards to the containing DAs. In Figure 9, the opinions **a bit more bouncy** and **a bit more durable**, along with their *same* target relation create the reinforcing opinion frame $< +, +, same >$. This relation is transferred upwards to the containing DAs, thereby creating a reinforcing relation between DA-2 and DA-4. Similarly, DA-4 and DA-5 are also in a reinforcing relation.

The data for our experiments consists of 7 scenario-based, multi-party meetings that were annotated using the procedure in Chapter 4. The annotations were produced by consensus. We filter out very small DAs (DAs with fewer than 3 tokens, punctuation included).[2] This gives us a total of 4606 DA instances, of which 1935 (42%) have opinion annotations.

5.2 FRAMEWORK FOR INCORPORATING DISCOURSE-LEVEL INFORMATION FOR POLARITY CLASSIFICATION

Our discourse-based analysis requires a global inference framework that takes into account the various local assignments and produces an optimal inference corresponding to the most coherent interpretation of the discourse. Our observations in previous chapters (recall the discussion on Example 1.2 in Chapter 3) was that when the opinion polarity of an expression is ambiguous, clear opinions in the surrounding discourse can help with the disambiguation. We design our experimental framework around this observation.

The polarity classification is designed as a two step process. In the first step, the classification is performed by looking at the words contained in the DA, that is, the local context

[2] Preliminary observations revealed that DAs with fewer than 3 tokens are less likely to contain opinion content – these are more often back-channels, interrupted utterances or utterances meant to establish common grounds. Thus naturally, these DAs would not participate in frame relations. Filtering these out helps us to reduce the class skew.

alone. This first classifier is called *Local*. The discourse information is used in the second step to create more optimal assignments that are coherent with the discourse. The classifier at this stage is called *Global*. Global bootstraps from the classifications produced by Local. It accommodates the preferences of the local classifier and for coherence with discourse neighbors.

This framework also enables us to test our hypothesis. The improvement of Global over Local will help us validate that discourse-level information is indeed useful over word-based methods.

We explore two very different paradigms for the global classifier. Specifically, we explore an optimization framework and a collective classification framework. These two approaches are very different in the way they incorporate the discourse information. The optimization framework employs Integer Linear Programming (ILP) and encodes discourse coherence as constraints that have to be satisfied. Using this framework, we are essentially encoding human prior knowledge of discourse coherence into the global inference system. This approach is unsupervised as there is no training involved. The collective classification framework, on the other hand, uses the annotated discourse relations as training data to learn a coherent interpretation of opinions in the discourse. Thus, this is a supervised framework requiring annotated data. In this work we use Iterative Collective Classification (ICA) which is an implementation of the collective classification algorithm. The ICA implementation is provided to us by Galileo Namata and Lise Getoor from University of Maryland at College Park (UMD).

5.3 LINGUISTIC FEATURES AND THE LOCAL CLASSIFIER

The local classifier, Local, is a supervised classifier using the SMO implementation of Weka [Hall et al., 2009]. It is important to make Local as reliable as possible; otherwise, the discourse relations will propagate misclassifications. Thus, we build Local using a variety of knowledge sources that have been shown to be useful for opinion analysis in previous work. Specifically, we construct features using unigrams, lexicons and dialog attributes.

We construct sentiment features using a sentiment lexicon that has been shown effective for sentiment analysis by [Wilson et al., 2005a]. This lexicon has about 8000 entries. Each entry is marked for its prior polarity (positive/negative/neutral).

We adapt an arguing lexicon from previous work [Somasundaran et al., 2007a] to create a polarity lexicon for the arguing category. The original arguing lexicon, created by Josef Ruppenhofer by inspecting the ICSI corpus, contains expressions (unigrams, bigrams and regular expressions) that indicate arguing. However, this lexicon does not have polarity information. We manually added the polarity information for each of the 226 entries.

Our work in [Somasundaran et al., 2007a] has explored the utility of Dialog Act (DA) tags for detecting the presence of opinions in sentences and utterances. Results from these experiments indicated that adding DA tag information to unigram information can improve the performance of sentiment and arguing detection. Insights from these experiments shaped our choice for features in this thesis. We reproduce some of the insights in Table 9 and Table 10.[3] Specifically, Table 9 reports the performance for arguing detection at the segment (utterance) and sentence level. We see that the Fmeasure is significantly improved when DA label information is encoded with unigram information into the supervised classifier (the results that are significant at $p < 0.05$ are in bold). Similarly, Table 10 shows that sentiment detection systems incorporating DA information can achieve significantly better performance than using unigrams alone. Note that our experiments in [Somasundaran et al., 2007a] are different from the experiments in this chapter. For example, the units of classification are sentences and segments, while here our units are DAs. Also, the classes predicted in [Somasundaran et al., 2007a] are subjectivity categories, while in this work we are interested in polarity classification. Nevertheless, we felt that the features discovered in those experiments would be useful for this work.

Using the resources discussed above, we construct a feature vector for each DA instance. For every input instance, the local classifier produces a three way classification of whether its polarity is positive, negative or neutral. It also outputs a class distribution for the input instance. The classifications are input for the ICA classifier, while the class distribution is input for the ILP classifier.

[3]Details of these experiments can be found in in [Somasundaran et al., 2007a].

	Acc	Prec	Rec	F-measure
	Segment Level classification			
Unigram	88.42	69.52	51.95	57.99
Unigram +DA	89.28	**73.81**	**54.62**	**61.26**
	Sentence Level classification			
Unigram	89.43	69.22	46.69	54.62
Unigram +DA	89.80	71.11	49.07	**56.7**

Table 9: Detection of arguing opinions at the sentence and the segment level is improved by using DA tag information

	Acc	Prec	Rec	F-measure
	Segment Level classification			
Unigram	86.87	80.84	48.53	58.77
Unigram +DA	87.45	81.93	**51.48**	**62.0**
	Sentence Level classification			
Unigram	88.23	82.41	44.08	56.61
Unigram +DA	88.59	82.11	**47.08**	**59.14**

Table 10: Detection of sentiment opinions at the sentence and the segment level is improved by using DA tag information

5.4 INTEGER LINEAR PROGRAMING

We now present the optimization paradigm which is implemented as an Integer Linear Programming problem. First, we discuss the intuition behind the ILP formulation. Then, we explain in detail how discourse constraints are actually encoded in the optimization problem.

Discourse relations between opinions can provide coherence constraints on the way their polarity is interpreted. While the local classifier's preference for a class should be considered, attention should also be paid to discourse coherence. Consider a discourse scenario in which a speaker expresses multiple opinions regarding the same thing, and is reinforcing his stance in the process (as in Example 1.2). The set of individual polarity assignments that is most coherent with this discourse is the one where all the opinions have the same (*equal*) polarity. On the other hand, a pair of individual polarity assignments most consistent with a discourse scenario where a speaker reinforces his stance via opinions towards alternative options, is one with opinions having mutually *opposite* polarity. For instance, in the utterance "Shapes **should** be *curved*, **nothing** *square-like*", the speaker reinforces his pro-curved shape stance via his opinions about the alternative shapes, *curved* and *square-like*. And, we see that the first opinion is positive and the second is negative. Table 11 lists the discourse relations (target and frame relation combinations) found in the corpus and the polarity interpretation for instances related by them.

Target relation + Opinion relation	Polarity
same+reinforcing	equal (e)
same+non-reinforcing	opposite (o)
alternative+reinforcing	opposite (o)
alternative+non-reinforcing	equal (e)

Table 11: Discourse relations and their polarity constraints on the related instances.

These constraints are encoded into a Binary Integer Linear Programming optimization method to impose discourse constraints on the classifications produced by Local. Here, the objective function is formulated to maximize the local prediction probabilities, while the

constraints are formulated to maintain discourse coherence.

5.4.1 Optimization Problem

For each DA instance i in a data set, the local classifier provides a class distribution $[p_i, q_i, r_i]$, where p_i, q_i and r_i correspond to the probabilities that i belongs to positive, negative and neutral categories, respectively. The optimization problem is formulated as an ILP minimization of the objective function in Equation 5.1.

$$-1 \times \sum_i (p_i x_i + q_i y_i + r_i z_i) + \sum_{i,j} \epsilon_{ij} + \sum_{i,j} \delta_{ij} \tag{5.1}$$

where the x_i, y_i and z_i are binary *class variables* corresponding to positive, negative and neutral classes, respectively. When a class variable is 1, the corresponding class is chosen. Variables ϵ_{ij} and δ_{ij} are binary *slack variables* that correspond to the discourse constraints between two distinct DA instances i and j. Specifically, ϵ_{ij} is used for producing slack in the equal polarity constraint between i and j. Similarly, δ_{ij} is used for producing slack in the opposite polarity constraint. When a given slack variable is 1, the corresponding discourse constraint is violated. Note that the objective function tries to achieve two goals. The first part $(\sum_i p_i x_i + q_i y_i + r_i z_i)$ is a maximization that tries to choose a classification for the instances that maximizes the probabilities provided by the local classifier. The second part $(\sum_{i,j} \epsilon_{ij} + \sum_{i,j} \delta_{ij})$ is a minimization that tries to minimize the number of slack variables used, that is, minimize the number of discourse constraints violated.

Constraints in Equations 5.2 and 5.3 listed below impose binary constraints on the variables. The constraint in Equation 5.4 ensures that, for each instance i, only one class variable is set to 1. This ensures exactly one class is assigned to each instance.

$$x_i \in \{0,1\}, y_i \in \{0,1\}, z_i \in \{0,1\} \ , \ \forall i \tag{5.2}$$

$$\epsilon_{ij} \in \{0,1\}, \delta_{ij} \in \{0,1\} \ , \ \forall i \neq j \tag{5.3}$$

$$x_i + y_i + z_i = 1 \ , \ \forall i \tag{5.4}$$

We pair distinct DA instances i and j as ij, and if there exists a discourse relation between them, they can be subject to the corresponding polarity constraints listed in Table

11. For this, we define two binary discourse-constraint constants: the *equal-polarity* constant, e_{ij} and the *opposite-polarity* constant, o_{ij}.

If a given DA pair ij is related by either a same+reinforcing relation or an alternative+non-reinforcing relation (rows 1, 4 of Table 11), then $e_{ij} = 1$; otherwise it is zero. Similarly, if it is related by either a same+non-reinforcing relation or an alternative+reinforcing relation (rows 2, 3 of Table 11), then $o_{ij} = 1$. Both e_{ij} and o_{ij} are zero if the instance pair is unrelated in the discourse.

For each DA instance pair ij, equal-polarity constraints are applied to the polarity variables of i (x_i, y_i) and j (x_j, y_j) via the following equations:

$$|x_i - x_j| \leq 1 - e_{ij} + \epsilon_{ij} \ , \ \forall i \neq j \tag{5.5}$$

$$|y_i - y_j| \leq 1 - e_{ij} + \epsilon_{ij} \ , \ \forall i \neq j \tag{5.6}$$

$$-(x_i + y_i) \leq -l_i \ , \quad \forall i \tag{5.7}$$

When $e_{ij} = 1$, the Equation 5.5 constrains x_i and x_j to be of the same value (both zero or both one) if there is no slack used. However, if the slack variable $\epsilon_{ij} = 1$, x_i and x_j can take any value independent of one other. That is, when $\epsilon_{ij} = 1$, the equal polarity discourse constraint between the instances i and j is violated. However, our objective function tries to minimize the slack used in the system.

Equation 5.6 similarly constrains y_i and y_j to be of the same value when $e_{ij} = 1$. Via these equations, we ensure that the instances i and j do not have the opposite polarity when $e_{ij} = 1$. However notice that, if we use just Equations 5.5 and 5.6, the optimization can converge to the same, non-polar (neutral) category. To guide the convergence to the same polar (positive or negative) category, we use Equation 5.7. Here $l_i = 1$ if the instance i participates in one or more discourse relations.

Next, the opposite-polarity constraints are applied via the following equations:

$$|x_i + x_j - 1| \leq 1 - o_{ij} + \delta_{ij} \ , \ \forall i \neq j \tag{5.8}$$

$$|y_i + y_j - 1| \leq 1 - o_{ij} + \delta_{ij} \ , \ \forall i \neq j \tag{5.9}$$

In the above equations, when $o_{ij} = 1$, x_i and x_j (and y_i and y_j) take on opposite values; for example, if $x_i = 1$ then $x_j = 0$ and vice versa. δ_{ij} is the slack variable that allows the opposite polarity discourse constraint between instances i and j to be violated. That is, when $\delta_{ij} = 1$, x_i and x_j (and y_i and y_j) take on any value independent of one another. However, we minimize the amount of discourse-constraint violations using the objective function. When $o_{ij} = 0$, the variable assignments are independent of one another.

In general, in our ILP formulation, notice that if an instance does not have a discourse relation to any other instance in the data, its classification is unaffected by the optimization. Also, as the underlying discourse scheme poses constraints only on the interpretation of the polarity of the related instances, discourse constraints are applied only to the polarity variables x and y, and not to the neutral class variable, z. Finally, even though slack variables are used, we discourage the ILP system from indiscriminately setting the slack variables to 1 by making them a part of the objective function that is minimized.

5.5 ITERATIVE COLLECTIVE CLASSIFICATION

In this section we present the collective classification paradigm. We use a variant of the Iterative Collective Classification Algorithm (ICA) [Lu and Getoor, 2003, Neville and Jensen, 2000], which is a collective classification algorithm shown to perform consistently well over a wide variety of relational data. It is important to note that the algorithm for ICA is implemented by our collaborators at UMD, Lise Getoor and Galileo Namata, and should not be considered a contribution of this thesis. We use ICA similarly to using a standard off-the-shelf machine learning algorithm. Our focus is the feature engineering to encode discourse-level information into the ICA framework for polarity classification.

The ICA uses two classifiers: a local classifier and a *relational classifier*. The local classifier is trained to predict the polarity of DA instances using only the local features. We use Local, described in Section 5.3, for this purpose. The relational classifier is trained using local features, and an additional set of features commonly referred to as *relational features*.

The value of a relational feature for a given DA depends on the polarity of the discourse

Percent of neighbors with polarity type a related via frame relation f'
Percent of neighbors with polarity type a related via target relation t'
Percent of neighbors with polarity type a related via frame relation f and target relation t
Percent of neighbors with polarity type a and same speaker related via frame relation f'
Percent of neighbors with polarity type a and same speaker related via target relation t'
Percent of neighbors with polarity type a related via a frame relation or target relation
Percent of neighbors with polarity type a related via a reinforcing frame relation or *same* target relation
Percent of neighbors with polarity type a related via a non-reinforcing frame relation or alt target relation
Most common polarity type of neighbors related via a *same* target relation
Most common polarity type of neighbors related via a reinforcing frame relation and *same* target relation

Table 12: Relational features: $a \in$ {non-neutral (i.e., positive or negative), positive, negative}, $t \in$ {same, alt}, $f \in$ {reinforcing, non-reinforcing}, $t' \in$ {same or alt, same, alt}, $f' \in$ {reinforcing or non-reinforcing, reinforcing, non-reinforcing}

for each instance i **do** {bootstrapping}
  Compute polarity for i using local attributes
end for
repeat {iterative}
  Generate ordering I over all instances
  for each i in I **do**
    Compute polarity for i using local and relational attributes
  end for
until Stopping criterion is met

Figure 10: The ICA Algorithm implemented by our UMD collaborators

neighbors of that DA. Relational features incorporate discourse and neighbor information; that is, they incorporate the information about the frame and target relations in conjunction with the polarity of discourse neighbors. Intuitively, our motivation for this approach can be explained using Example 1.2 (reproduced below).

(1.2) D:: ... this kind of rubbery material, *it's* a **bit more bouncy**, like you said they get chucked around a lot. A **bit more durable** and *that* can also be **ergonomic** and *it* kind of feels **a bit different from all the other remote controls**.

Here, in interpreting the ambiguous opinion **a bit different** as being positive, we use the knowledge that it participates in a reinforcing discourse, and that all its neighbors (e.g., **ergonomic**, **durable**) are positive opinions regarding the same thing. On the other hand, if it had been a non-reinforcing discourse, then the polarity of **a bit different**, when viewed with respect to the other opinions, could have been interpreted as negative.

Table 22 lists the relational features we defined for our experiments where each row represents a set of features. Features are generated for all combinations of a, t, t', f and f' for each row. For example, one of the features in the first row is *Percent of neighbors with polarity type positive, that are related via a reinforcing frame relation.* Thus, each feature

for the relational classifier identifies neighbors for a given instance via a specific relation (f, t, f' or t', obtained from the scheme annotations) and factors in their polarity values (a, obtained from the classifier predictions from the previous round). This adds a total of 59 relational features to the already existing local features.

The pseudocode for the ICA algorithm implemented by the UMD team is shown in Figure 10. ICA has two main phases: the bootstrapping and iterative phases. In the bootstrapping phase, the polarity of each instance is initialized to the most likely value given only the local classifier and its features. In the iterative phase, random ordering of all the instances is created and, the relational classifier is applied, in turn, to each instance where the relational features, for a given instance, are computed using the most recent polarity assignments of its neighbors. This is repeated until some stopping criterion is met. For our experiments, this was a fixed number of 30 iterations, which has been found to be sufficient in most data sets for ICA to converge to a solution [Sen et al., 2008] (in fact, most ICA-based algorithms have been found to converge within 10 iterations [McDowell et al., 2009]).

5.6 EXPERIMENTS

In this work, we are particularly interested in improvements due to discourse-level relations. Thus, we report performance under three conditions: over only those instances that are related via discourse relations (*Connected*), over instances not related via discourse relations (*Singletons*), and over all instances (*All*). It is expected that the discourse-based methods will create a greater impact for the Connected condition. Table 13 shows the class distributions in the data for the three conditions.

5.6.1 Classifiers

Our first baseline, *Base*, is a simple distribution-based classifier that classifies the test data based on the overall distribution of the classes in the training data. However, in Table 13, the class distributions are different for the Connected and Singleton conditions. We incorporate

	Pos	Neg	Neutral	Total
Connected	643	343	81	1067
Singleton	553	233	2753	3539
All	1196	576	2834	4606

Table 13: Class distribution over connected, single and all instances.

this in a smarter baseline, *Base-2*, which constructs separate distributions for connected instances and singletons. Thus, given a test instance, depending on whether it is connected, Base-2 uses the corresponding distribution to make its prediction.

The third baseline is the supervised classifier, Local, described in Section 5.3. It is implemented using the SVM classifiers from the Weka toolkit [Witten and Frank, 2002]. We use the SMO implementation, which, when used with logistic regression, has an output that can be viewed as a posterior probability distribution. The supervised discourse-based classifier, ICA from Section 5.5, also uses a similar SVM implementation for its relational classifier. We implement our ILP approach from Section 5.4 using the optimization toolbox from Mathworks (http://www.mathworks.com) and GNU Linear Programming Kit.

We observed that the ILP system performs better than the ICA system on instances that are connected, while ICA performs better on singletons. Thus, we also implemented a simple hybrid classifier (HYB), which selects the ICA prediction for classification of singletons and the ILP prediction for classification of connected instances.

5.6.2 Results

We performed 7-fold cross validation experiments, where six meetings are used for training and the seventh is used for testing the supervised classifiers (Base, Base-2, Local and ICA). In the case of ILP, the optimization is applied to the output of Local for each test fold. Table 14 reports the accuracies of the classifiers, averaged over 7 folds. The accuracies are calculated as follows:

	Base	Base-2	Local	ICA	ILP	HYB
Connected	24.4	47.56	46.66	**55.64**	**75.07**	**75.07**
Singleton	51.72	63.23	75.73	<u>78.72</u>	75.73	<u>78.72</u>
All	45.34	59.46	68.72	**73.31**	**75.35**	**77.72**

Table 14: Accuracies of the classifiers measured over Connected, Singleton and All instances. Performance significantly better than Local are indicated in **bold** for $p < 0.001$ and <u>underline</u> for $p < 0.01$.

$$Accuracy_{connected} = \frac{\#\ correct\ guesses\ for\ connected\ instances}{total\ \#\ connected\ instances}$$

$$Accuracy_{singleton} = \frac{\#\ correct\ guesses\ for\ singleton\ instances}{total\ \#\ singleton\ instances}$$

$$Accuracy_{all} = \frac{\#\ correct\ guesses}{total\ \#\ instances}$$

First, we observe that Base performs poorly over connected instances, but performs considerably better over singletons. This is expected as the overall majority class is neutral and the singletons are more likely to be neutral. Base-2, which incorporates the differentiated distributions, performs substantially better than Base.

Local achieves an overall performance improvement over Base and Base-2 by 23 percentage points and 9 percentage points, respectively. In general, Local outperforms Base for all three conditions ($p < 0.001$), and Base-2 for the Singleton and All conditions ($p < 0.001$). This overall improvement in Local's accuracy corroborates the utility of the lexical, unigram and DA-based features for polarity detection in this corpus.

Turning to the discourse-based classifiers, ICA, ILP and HYB, all of these perform better than Base and Base-2 for all conditions. ICA improves over Local by 9 percentage points for Connected, 3 points for Singleton and 4 points for All. ILP's improvement over Local for Connected and All is even more substantial: 28 percentage points and 6 points, respectively. Notice that ILP has the same performance as Local for Singletons as there are no discourse constraints for unconnected instances. Finally, HYB significantly outperforms Local under

all conditions. The significance levels of the improvements over Local are highlighted in Table 14. These improvements also signify that the underlying discourse scheme is effective, and adaptable to different implementations.

Interestingly, ICA and ILP improve over Local in different ways. While ILP sharply improves the performance over the connected instances, ICA shows relatively modest improvements over both Connected and Singletons. ICA's improvement over Singletons is interesting because it indicates that, even though the features in Table 22 are focused on discourse relations, ICA utilizes them to learn the classification of singletons too.

Comparing our discourse-based approaches, ILP does significantly better than ICA over connected instances ($p < 0.001$), while ICA does significantly better than ILP over singletons ($p < 0.01$). However, there is no significant difference between ICA and ILP for the All condition. The HYB classifier outperforms ILP for the Singleton condition ($p < 0.01$) and ICA for the Connected condition ($p < 0.001$). Interestingly, over all instances (the All condition), HYB also performs significantly better than *both* ICA ($p < 0.001$) and ILP ($p < 0.01$).

5.6.3 Analysis

We measured the precision, recall and Fmeasure of the best performing baseline (Local) and our two discourse-based systems, ICA and ILP, for the three classes (positive, negative, neutral) and three conditions (Connected, Singletons and Overall). Note that, as HYB is same as ILP for Connected, and same as ICA for Singletons, we do not present an analysis for HYB. Tables 15, 16 and 17 report the performance. All numbers reported are averaged over 7 fold cross-validation experiments. The metrics precision, recall and Fmeasure have their standard definitions as used in information retrieval. For example,

$$Precision_{positive} = \frac{\#\ correct\ guesses\ for\ the\ positive\ class}{total\ \#\ guesses\ made\ as\ positive\ class}$$

$$Recall_{positive} = \frac{\#\ correct\ guesses\ for\ the\ positive\ class}{total\ \#\ positive\ instances\ in\ the\ data}$$

$$Fmeasure_{positive} = \frac{2 * Precision_{positive} * Recall_{positive}}{Precision_{positive} + Recall_{positive}}$$

Polarity Class	Metric	Local	ICA	ILP
Positive	Precision	78.1	78.0	78.2
	Recall	45.3	55.0	86.3
	Fmeasure	56.8	64.3	81.5
Negative	Precision	71.9	61.3	69.8
	Recall	44.1	57.8	73.4
	Fmeasure	54.0	59.4	70.7
Neutral	Precision	12.1	11.7	
	Recall	62.8	41.0	*
	Fmeasure	18.5	16.4	

Table 15: Precision, Recall, Fmeasure for the connected instances

For the connected instances (Table 15) we observe that ICA and ILP both have substantially better recall than Local for the polar classes. The improvement by ICA is 10 percentage points for the positive class and 13 percentage points for the negative class. ILP produces more substantial improvements – as compared to Local, ILP has a 29 percentage point improvement for the negative class and a 41 percentage point improvement for the positive class.

The improvements in recall in both systems is accompanied by a small drop in precision. ILP's precision stays the same for the positive class, while it drops by 2 percentage points for the negative class. ICA's precision remains the same for the positive class, but there is a drop in precision for the negative class by about 10 percentage points. However, the gains in recall for each of the polar categories offset the corresponding drops in precision for both ICA and ILP. This is seen in the form of improved Fmeasures for ICA for both, the positive (8 percentage points) and negative (5 percentage points) classes, as well as for ILP for both, the positive (25 percentage points) and negative (16 percentage points) classes.

For the neutral instances, however, the behavior is different. By virtue of the constraint in Equation 5.7, ILP does not classify any connected instance as neutral; thus the precision

is undefined, the recall is 0 and the Fmeasure is undefined. This is indicated as * in Table 15. There is a drop in the performance over all metrics when ICA makes predictions for connected instances belonging to the neutral class. We believe this is due to the fact that for the Connected condition, there are very few instances that are neutral (see row 1 of distribution table Table 13) which makes it difficult for ICA to learn the classification of these instances.

Additionally, the poor performance of Local and ICA for neutral Connected instances can be explained by the fact that connected DAs are neutral only because of the random polarity assignment, which happens when there are multiple opinions in a DA, picked the neutral opinion instead of the polar opinion. In such cases, there are features in the DA that are indicative of the polar class, which eventually cause wrong classifications.

Polarity Class	Metric	Local	ICA	ILP
Positive	Precision	43.7	48.9	43.7
	Recall	47.9	35.2	47.9
	Fmeasure	44.9	40.3	44.9
Negative	Precision	38.6	45.0	38.6
	Recall	45.2	29.0	45.2
	Fmeasure	39.7	33.0	39.7
Neutral	Precision	88.3	85.0	88.3
	Recall	84.3	92.0	84.3
	Fmeasure	86.1	88.3	86.1

Table 16: Precision, Recall, Fmeasure for the Singleton instances

Table 16 reports the performance of all three systems over singletons. Observe that ILP performs no different than Local in this table as the discourse constraints are not applied over unconnected instances. ICA shows an improvement in precision for both polar classes (5 percentage points for the positive class and 6 percentage points for the negative class). However the drop in recall is relatively higher (12 percentage point drop for positive class and 16 percentage point drop for negative class), which results in an overall drop in the

Fmeasure for both polar classes. However, ICA shows an improvement in recall and as a consequence, Fmeasure, for the neutral instances. This performance is seen due to the fact that ICA makes very conservative polar guesses for singletons.

The performance of ICA in Tables 15 and 16 indicate that it does well for the polar categories for connected instances and the neutral category for singletons, but performs poorly for the polar categories for singletons and the neutral category for connected instances. This behavior may be due to the fact that for the connected instances, there are very few neutral cases, which makes learning difficult for the supervised learner. On the other hand, in the case of singletons, a large skew in the data (about 78% of singletons are neutral) causes a bias in the learner.

Finally, Table 17 reports the performance of all three systems over all instances in the data. Here we observe that ILP improves over Local over all metrics for both positive and negative classes. ICA also improves over Local for precision and Fmeasure (though there is a fractional drop in the recall for the positive category). Also, both ILP and ICA show an overall performance improvement in Fmeasure for the neutral category. Thus, for the Fmeasure metric computed over the entire data the discourse-based approaches show improvement for all polar classes over Local.

Polarity class	Metric	Local	ICA	ILP
Positive	Precision	56.2	64.5	61.3
	Recall	46.6	46.0	67.7
	Fmeasure	50.4	53.5	64.0
Negative	Precision	52.3	54.6	54.6
	Recall	44.3	46.3	62.5
	Fmeasure	46.0	49.4	57.1
Neutral	Precision	76.3	78.4	88.3
	Recall	83.9	90.5	81.5
	Fmeasure	79.6	83.9	84.6

Table 17: Precision, Recall, Fmeasure for All instances

Gold	Local			
	Pos	Neg	Neut	Total
Pos	551	113	532	1196
Neg	121	250	205	576
Neut	312	135	2387	2834
Total	984	498	3124	4606

Gold	ILP			
	Pos	Neg	Neut	Total
Pos	817	157	222	1196
Neg	147	358	71	576
Neut	358	147	2329	2834
Total	1322	662	2622	4606

Table 18: Contingency table over all instances.

Notice that, looking at performance figures over Tables 15, 16 and 17, some numbers may seem counter-intuitive. For example, ILP has an *overall improvement in precision over Local*, even though for the connected condition (Table 15) it does similar to, or lower than Local. This is explained by the fact that, while going from connected to overall conditions, Local's polar predictions increase by threefold (565 to 1482), but its *correct* polar predictions increase by only twofold (430 to 801). Thus, the ratio of change in the total polar predictions to the correct polar predictions is 3 : 2. On the other hand, while polar predictions by ILP increase by only twofold (1067 to 1984), its *correct* polar predictions increase by 1.5 times (804 to 1175). Here, the ratio of change in the total polar predictions to the correct polar predictions is 4 : 3, a smaller ratio. The contingency table (Table 18) explains this. It shows how Local and ILP compare against the gold standard annotations. Notice here, that even though ILP makes more polar guesses as compared to Local, a greater proportion of the ILP guesses are correct.

5.7 DISCUSSION

In this chapter, we measure the impact of employing discourse-based global information over the local word-based method. Our discourse information is comprised of three different pieces of information: target relations, opinion relations and polarity of the neighboring instances. Of these three pieces, we obtain the first two from our manual annotations; hence they are reliable. However, we obtain the polarity of the neighbors from the local classifier (or the previous iteration in the case of ICA). Observe in Table 14 that the accuracy of Local is less than 50% for the connected instances. This indicates that the connected instances are inherently complex. Local provides the classifications for bootstrapping ICA and ILP. Methods starting with noisy starting points are in danger of propagating the errors and hence worsening the performance. Interestingly, in spite of starting with so many bad classifications, ILP and ICA are able to achieve performance improvements. In this section we look at some examples and compare the performance of Local versus our best performing classifier, ILP.

We discovered that, given a set of connected instances, even when Local has only one correct guess, ILP is able to use this to rectify the related instances. We illustrate this situation in Figure 11, which reproduces the connected DAs for Example 1.2. It shows the classifications for each DA from the gold standard (in green boxes), the Local classifier (L, in grey boxes) and the ILP classifier (ILP, in blue boxes). Observe that Local predicts the correct class (positive) for only DA-4 (the DA containing **bit more durable** and **ergonomic**). These are clear cases of positive evaluation. It incorrectly predicts the polarity of DA-2 (containing **bit more bouncy**) as neutral (*), and DA-5 (containing **a bit different from all the other remote controls**) as negative (-). DA-2 and DA-5 exemplify the fact that polarity classification is a complex and difficult problem: being bouncy is a positive evaluation in this particular discourse context, and may not be so elsewhere. Thus, naturally, lexicons and unigram-based learning would fail to capture this positive evaluation.

Similarly, "being different" could be deemed negative in other discourse contexts. However, ILP is able to arrive at the correct predictions for all the instances. As DA-4 is connected to both DA-2 and DA-5 via discourse relations that enforce an equal-polarity

constraint (same+reinforcing relation of row 1, Table 11), both misclassifications are rectified. Presumably, the probability mass assigned by Local for the instances with incorrect predictions was not very biased towards the wrong class, but the probability mass was highly biased towards the correct class for the correct guess. Thus, in the process of maintaining discourse constraints, ILP was able to change the polarity of the cases about which Local was not confident, thereby resulting in the correct classifications.

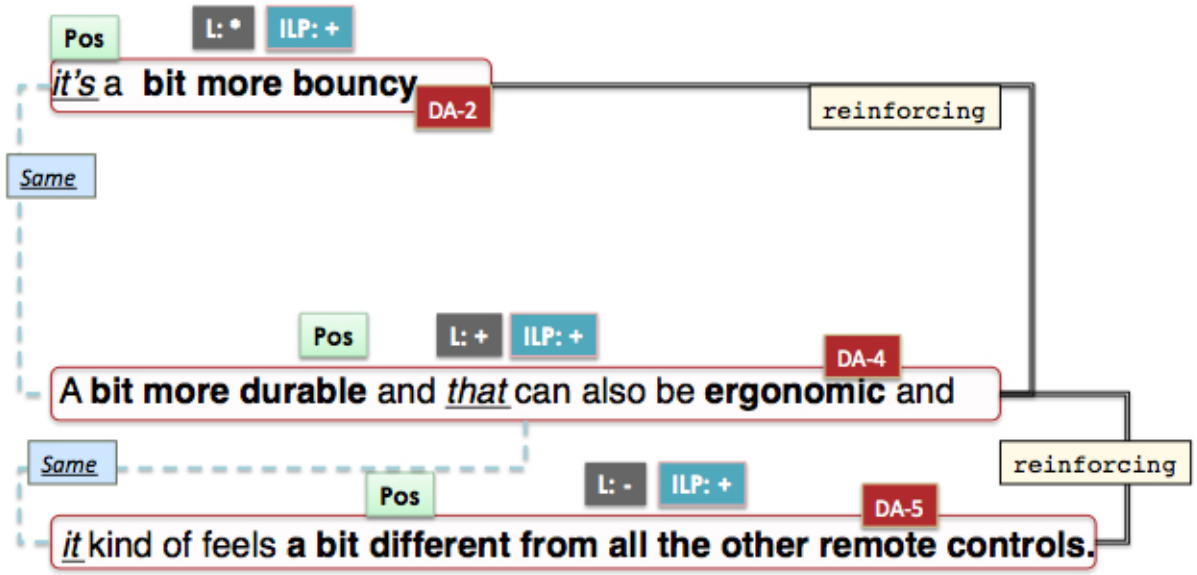

Figure 11: Gold standard classification and classifications produced by Local and ILP Example 1.2

We also observed the propagation of correct classifications for other types of discourse relations, for more complex types of connectivity, and also for conditions where an instance is not directly connected to a correctly predicted instance. The meeting snippet below (Example 5.10) and its corresponding DA relations (Figure 12) illustrate this. This example is a reinforcing discourse where the speaker is arguing for the number keypad, which is an alternative to the scrolling option. Thus, he argues against the scrolling, and argues for entering the number (which is a capability of the number keypad).

(5.10) D-1: I reckon you're **gonna have to have** a *number keypad* anyway for the amount of channels these days,

D-2: You **wouldn't want to** just have to *scroll through* all the channels to get to the one you want

D-3: You **wanna** *enter just the number of it* , if you know it

D-4: I reckon **we're gonna have to have** a *number keypad* anyway

In Figure 12, we see that, of the four instances, Local predicts only DA-4 correctly. This instance is in a same+reinforcing relation with DA-3, which influences ILP to select a positive classification for DA-3. Next, DA-2 is connected via an alternative+reinforcing discourse relation to each of its neighbors, DA-1 and DA-3, which encourages the optimization to choose a class for it that is opposite to DA-1 and DA-3. Adding to this is the same+reinforcing relation between DA-1 and DA-3, driving the optimization to choose like polarities for them. Notice that even though Local predicts only DA-4 correctly, this correct classification finally influences the correct choice for all the instances, including the remotely connected DA-1 and DA-2.

5.8 RELATED WORK

Hatzivassiloglou and McKeown [Hatzivassiloglou and McKeown, 1997] use conjunctions to constrain polarity interpretation of adjectives. For example, two adjectives conjoined with an "and" are likely to be of the same polarity. Our work is not limited to a particular part of speech tag, is not dependent on the presence of conjunctions and employs discourse-level relations on text spans that are not necessarily adjacent.

There has been a lot of effort in opinion research for finding the polarity of words using relationships between words in a corpus [Turney, 2002, Turney and Littman, 2003, Gamon and Aue, 2005] or in lexical resources (e.g. WordNet) [Takamura et al., 2005, Esuli and Sebastiani, 2005, Gyamfi et al., 2009, Andreevskaia and Bergler, 2006, Mohammad et al., 2009, Kamps et al., 2004, Kim and Hovy, 2004]. These works focus on prior polarity. Determining semantic orientation of words out of context is useful to build polarity lexicons. Our work is complementary to and builds on this word-based information. As seen in Example 1.2, clear cases of opinions (e.g., durable, ergonomic) can

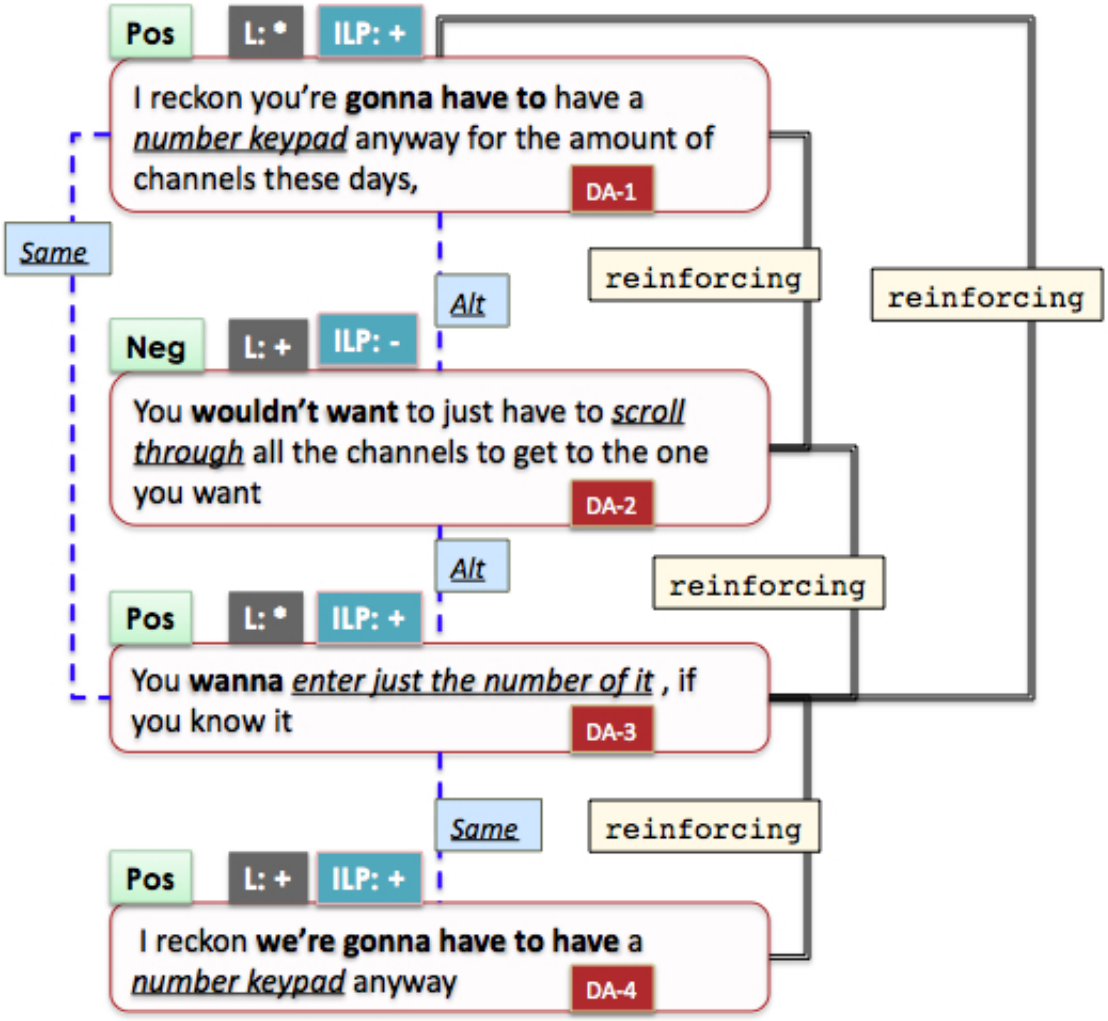

Figure 12: Gold standard classification and classifications produced by Local and ILP Example 5.10

be resolved using a polarity lexicon. We employ our discourse-based approach to use these resolved cases to disambiguate difficult instances.

Researchers have also used contextual clues to disambiguate subjectivity [Riloff and Wiebe, 2003] and polarity [Yi et al., 2003a, Suzuki et al., 2006, Wilson et al., 2005a, Kennedy and Inkpen, 2005, Kanayama and Nasukawa, 2006, Devitt and Ahmad, 2007]. Our work focuses on using the surrounding discourse, particularly, the *opinions* in the surrounding discourse for disambiguating polarity. There is previous work on the effect of reversal words such as "but" on polarity [Sadamitsu et al., 2008]. Here, when a reversal word is encountered, the current (prior) polarity is flipped. However, they do not capture discourse-level relations.

Researchers, such as [Polanyi and Zaenen, 2006], have discussed how the discourse structure can influence opinion interpretation; and previous work, such as [Asher et al., 2008], have developed annotation schemes for interpreting opinions with discourse relations. However, they do not empirically demonstrate how automatic methods can use their ideas to improve polarity classification. In this work, we demonstrate concrete ways in which a discourse-based scheme can be modeled using global inference paradigms.

Using min-cut on sentential graph has been a popular global approach to take into account cohesion between sentences [Pang et al., 2002, Pang and Lee, 2004]. Min-cut is a natural formulation when nodes in a graph have binary relationships: either they cohere or they do not. This framework encodes the human prior knowledge that sentences that have high cohesion (where cohesion can be determined by adjacency, cosine measure, amount of positive word usage etc.) should have the same polarity tag. In our work, we take the ILP approach as we have 3 types of relationships: relations that influence nodes to have same polarity (which can be considered similar to cohesion), no relation (which can be considered similar to non-cohesion) and relations that influence nodes to have opposite polarity. Another difference between this work and the above works that use min-cut is that we use our global inference to resolve fine-grained polarity, while theirs is a document-level classification problem.

Popescu and Etzioni [Popescu and Etzioni, 2005] use relaxation labeling, a collective classification approach, to find contextual polarity of words that are heads of opinion phrases.

However, their relationships between words are defined based on morphosyntactic similarities.

The biggest difference between this work and previous work in opinion analysis that use global inference methods is in the type of linguistic relations used to achieve the global inference. Some of the work is not related to discourse at all (e.g., lexical similarities [Takamura et al., 2007], word-based measures like TF-IDF [Goldberg and Zhu, 2006], agreement/disagreement between speakers [Thomas et al., 2006, Bansal et al., 2008], agreement between low-level tags [Snyder and Barzilay, 2007] or structural adjacency [McDonald et al., 2007]). In contrast, our work focuses on discourse-based relations for global inference. Another difference is that our work in this chapter is carried out over multi-party conversational data.

Researchers have previously used global inferences in meeting data to solve different problems such as detecting agreement/disagreement [Galley et al., 2004, Germesin and Wilson, 2009, Hillard et al., 2003]. Previous work on emotion and subjectivity detection in multi-party conversations has explored using prosodic information [Neiberg et al., 2006], combining linguistic and acoustic information [Raaijmakers et al., 2008] and combining lexical and dialog information [Somasundaran et al., 2007a]. This work is focused on harnessing discourse-based knowledge and on interdependent inference.

Joint models have been previously explored for other NLP problems [Haghighi et al., 2005, Moschitti et al., 2006, Moschitti, 2009]. Our global inference model focuses on opinion polarity recognition task. There are several collective classification frameworks, including [Neville and Jensen, 2000, Lu and Getoor, 2003, Taskar et al., 2004, Richardson and Domingos, 2006, Bilgic et al., 2007]. In this paper, we use an approach by [Lu and Getoor, 2003] which iteratively predicts class values using local and relational features. ILP has been used on other NLP tasks, e.g., [Denis and Baldridge, 2007, Choi et al., 2006, Carvalho and Cohen, 2005, Roth and Yih, 2004]. In this work, we employ ILP for modeling discourse constraints for polarity classification.

5.9 SUMMARY

In this chapter we demonstrated empirically how discourse-level relations can be employed for performing fine-grained opinion analysis. We first created a word-based polarity classifier that employs the state-of-the-art word-based and local information in a standard supervised classification framework. Specifically, we used unigram and sentiment lexicon information, resources that have been proven effective in previous work. Additionally, we adapted an arguing lexicon from previous work and also employed DA tag information. The resulting local classifier achieves an overall performance better than a smart distribution-based approach. Building this strong local classifier satisfied two objectives: First, this created a reliable classifier from which to bootstrap the discourse-based classifiers. Second, this gave us a strong word-based baseline for our empirical studies.

We explored different global paradigms to encode the discourse-level information. Specifically, we explored a supervised collective classification approach and an unsupervised optimization approach. The ILP approach encodes human prior knowledge of the discourse. It does not require training data, but does require the human knowledge to be encoded as constraints. ICA is an alternative paradigm, where no prior knowledge of the discourse is needed. Instead, training data is used to learn the discourse-based interdependent interpretation of polarity in a supervised fashion. The ICA approach is akin to relaxation labeling used by previous sentiment researchers [Popescu and Etzioni, 2005] and the ILP approach shares similarities with min-cut for document-level sentiment classification [Pang et al., 2002, Pang and Lee, 2004, Thomas et al., 2006].

Both our discourse-based approaches perform better than the word-based approach, which support both hypotheses 5.a and 5.b. Our results also provide evidence for the overall general hypothesis that the discourse-level relations in this thesis are useful for fine-grained opinion analysis. The improvements obtained by our diverse global inference approaches also indicate that discourse information can be adapted in different ways to augment and improve existing opinion analysis techniques.

We discovered that each of our approaches achieves performance improvements differently. The ICA approach shows accuracy improvements for singletons as well as connected

instances, while the ILP approach produces pronounced improvements only for the connected instances. We found that by combining the two approaches in a simple hybrid approach we can achieve the best overall accuracy. In our current set of experiments, ILP emerged an overall better and more consistent performer of the two discourse-based methods. We found that, in conditions where there is not enough training data (for e.g. the neutral category under the connected condition), or there is a large class skew (for e.g. the neutral category under the singleton conditions), ICA is unable to produce correct classifications. We found the that the ILP approach is able to achieve a large increase in recall of the polar categories without harming the precision, which results in its performance improvements.

Qualitatively, we illustrated how, even if the bootstrapping process is noisy, the optimization and discourse constraints effectively rectify misclassifications. We observed that even if the local classifier got a fourth of the instances correct, the discourse-based global inference was able to arrive at the correct classifications for all instances.

All experiments in this section are performed using perfect discourse relation information. Recognition of discourse-based relations is a challenging problem. The behavior of ICA and ILP can change, depending on the automation of discourse-level recognition. For example, when the discourse-based relationships are noisy, a supervised system such as ICA may achieve better relative improvements. In the next chapter we will attempt to learn the discourse relations automatically and employ these for polarity classification.

6.0 LEARNING DISCOURSE-LEVEL RELATIONS FROM ANNOTATIONS

In the previous chapter, we saw the usefulness of discourse-level relations for fine-grained opinion analysis. In this chapter we test the high-level hypothesis that *automatic systems can be developed for recognizing discourse-level relations better than baseline methods.*

We learn discourse-level relations from our annotated corpora using a supervised learning framework. First, we perform a preliminary study, where we explore a simplified problem of whether we can *detect the presence* of discourse-level opinion relations between opinion-bearing sentences. This is a binary classification task (presence/absence), where the input is a sentence pair, and the output is the binary classification.

Next, we will attempt to recognize each of the discourse-level relations, that is, the target and opinion relations (note that, as opinion relations exist due to opinion frames, we use the terms "frame relations" and "opinion relations" interchangeably in this work). Recognition of target relations involves detecting the presence of target relations and classifying their type, whether they are *same* or *alternative*. This is done as a three-way classification where the classes are *no-link*, *same* and *alternative*. We construct a target link classifier, which takes pairs of instances (sentences or other units such as DA), and produces the three-way classification. Similarly, recognition of opinion frame relations involves the following three-way classification: *no-link*, *reinforcing* and *non-reinforcing*. For recognizing opinion frame relations, we construct a frame link classifier, which takes pairs of instances and outputs the three-way classification.

The main goal of this chapter is to learn discourse-level relations from annotations. Under this umbrella, we have two specific goals:

Finding linguistic clues that help in the recognition of discourse-level relations:

First we investigate linguistically motivated clues that can help the detection of relations between opinion-bearing sentences. We explore clues based on discourse continuity, coreference relations, PDTB relations [Miltsakaki et al., 2004] and dialog-level relations.

Design and explore a global inference paradigm for recognizing discourse-level relations: Linguistic clues capture information in the local context. We have observed in Section 3.3 that an interdependent interpretation can also disambiguate discourse-level relations, if the global polarity information is clear. In order to investigate this idea further, we design a global inference paradigm where opinions, target relations and frame relations are classified in an inter-dependent fashion.

These goals provide us with two specific hypotheses under our main hypothesis:

Hypothesis 6.a Discourse-relation classifiers constructed using a variety of linguistic features perform better than baseline methods.

Hypothesis 6.b Discourse-relation classifiers constructed using global discourse features perform better than local baseline methods.

The rest of this chapter is organized as follows: First, we present our preliminary experiments in Section 6.1. Discourse-level relation recognition is presented in Section 6.2. We discuss the challenges and future directions in Section 6.3, discuss related work in Section 6.4 and conclude in Section 6.5.

6.1 PRELIMINARY STUDY: DETECTING OPINION FRAME RELATIONS

In our preliminary study, we explore a simplified problem: we only have to detect the presence of an opinion frame relation. We do not attempt to determine the exact nature of this relationship (whether the relationship is reinforcing or non-reinforcing). Also, we do not attempt to find relations between targets explicitly, even though target relations are an integral part of creating opinion frame annotations. The intention is to capture the relations

implicitly using machine learning features that encode this information. Our experiments focus on the question: What linguistic features are useful for detecting discourse-level opinion frame relations?

6.1.1 Data

The data for experiments in this study consist of 5 AMI meetings. This corpus is comprised of 4436 sentences or 2942 segments (utterances). The annotations (polarity labels and relations) are transferred upwards, to the containing sentences, similarly to our approach in the previous chapter. The unit of classification is a *sentence pair*.

We filter out neutral sentences and sentences that are less than two words in length. This filters out very small sentences (e.g., "Cool.") which rarely contain opinions that participate in frame relations. The length restriction is imposed to handle data skewness. Neutral sentences, by definition, will not have opinion relations and hence we filter these out from the preliminary analysis. This further mitigates the data skewness problem. Thus, our experimental data consists of pairs of opinion sentences and the gold-standard information whether there exists an opinion frame relation between them. We approximate continuous discourse by only pairing sentences that are not more than n sentences apart, where n was arbitrarily set to 10. Also, in order to alleviate data skewness, we only pair sentences belonging to the same speaker (sentences belonging to the same speaker are more likely to be related via frames than sentences belonging to different speakers). The experiments were performed on a total of 2539 sentence pairs, of which 551 are positive instances. Notice that this data is still quite skewed – only 21% of the instances are positive examples.

6.1.2 Features

The factor that determines if two opinions are related is primarily the target relation between them. Instead of first finding the target span for each opinion sentence and then inferring if they should be related, we directly try to encode target relation information in our features. We explored a number of features to incorporate this. The set that gives the best performance are listed in Table 19. These features are intended to capture discourse and dialog continuity

Time difference between the sentence pair
Number of intervening sentences
Content word overlap between the sentence pair
Anaphoric indicator in the second sentence
Existence of adjacency pair between the sentence pair
Focus space overlap between the sentence pair
Bag of words for each sentence

Table 19: Features for opinion relation detection

as well as target relations.

The *time difference* between the sentences and the *number of intervening sentences* are useful features to capture the idea that topics shift with time. Together, they capture discourse continuity. If two sentences are close to each other in time, it is likely that they belong to the same discourse context, thereby increasing the likelihood that the opinions in them are related. The number of intervening sentences also captures the idea of discourse continuity – if there are a large number of intervening sentences between a sentence pair, it is likely that the two sentences in the pair belong to different discourse contexts. In multiparty conversations, a large time difference between sentences does not always translate to a large number of intervening sentences. In meeting "hot-spots" (heated discussions), many participants speak simultaneously within a very short duration. Thus, there are many intervening sentences between a sentence pair, but they do not have a large time difference. On the other hand, if there is a pause (as is the case when participants pause to articulate their thoughts) two sentences may be nearby, but have a time gap between them.

The *content word overlap* feature captures the degree of topic overlap between the sentence pair, and looks for target relations via identity. This feature is intended to capture the relatedness between two sentences that contain, for example,

the words "square corners" and "square like". Content words are obtained by filtering out stop words from the sentence (we use a standard stop word list from http://ftp.dcs.glasgow.ac.uk/idom/ir_resources/linguistic_utils/stop_words). The *anaphoric indicator feature* checks for the presence of pronouns such as *it* and *that* in the second sentence (the second sentence is the sentence that occurs later in time) to account for target relations via anaphora. These two features essentially capture coreference-based target relations between the two opinion sentences.

The *focus space overlap* feature is motivated by our observation that participants refer to an established discourse topic without explicitly referring to it. This feature is inspired by the idea of focus from Grosz and Sidner's work on discourse structure. In an ongoing discussion, participants do not refer to the topic in every successive sentence. The evoked (and established) topic can be considered active in the discourse, until there is a an explicit topic shift, or a gradual topic drift. The focus space is designed to capture this phenomenon. A focus space is a finite sized list containing the most recently used noun phrases (NPs) in the discourse (a standard part of speech tagger is used for this purpose). A focus space is constructed for each sentence and consists of all the NPs in that sentence and the previous sentences. Thus, this structure holds all the NPs that are relevant to the current sentence, that is, the discourse focus. The corresponding feature for a given sentence pair is the percent overlap between the focus spaces of the two sentences. If a sentence does not introduce any new topics, it will have the same focus space as the previous sentence. Also, even if two sentences are separated by intervening sentences and separated in time, but no new topic is introduced, the sentences will still have a significant focus space overlap. On the other hand, if two sentences are adjacent, but the second sentence introduces a new topic for discussion, the sentences will not have a significant focus overlap. This feature is thus complementary to other features that capture time-based adjacency, distance-based adjacency or explicit target mentions.

The *existence of an adjacency pair* between the sentences can clue the system that the opinions in the sentences are related too. Adjacency pairs are manual dialog annotations available in the AMI corpus. Adjacency pair annotations link text spans that are related via dialog intentions, and these dialog-level relations might be indicative of opinion relations.

	Acc.	Prec.	Recall	Fmeasure
False	78.3%	-	0%	-
Distribution	66%	21.7%	21.7%	21.4%
Random	50.0%	21.5%	49.4%	29.8 %
True	21.7%	21.6%	100%	35.5 %
System	67.6%	36.8%	64.9%	46%

Table 20: Automatic detection of opinion frames relations

For example, if the second opinion sentence is an assessment of the first, it is likely that the opinions contained in them are also related via opinion frame relations.

Finally, unigrams, encoded as standard *bag of words* features, are included for each sentence.

6.1.3 Results

We performed 5-fold cross validation experiments, using 4 meetings to train and the remaining one to test in each fold. The results are averaged over the 5 folds. For machine learning, we use the standard SVMperf package [Joachims, 2005], an implementation of SVMs designed for optimizing multivariate performance measures. We found that, on our skewed data, optimizing on Fmeasure obtains the best results.

Our system is compared to four baselines in Table 20. The metrics here are overall accuracy, and precision, recall and Fmeasure for the prediction of the class "link present". The majority class baseline, which always guesses false (*False*) has good accuracy due to the data skew, but zero recall. As it makes no guess for "link present" its precision, and consequently Fmeasure is undefined. The baseline that always guesses true (*True*) has 100% recall and the best Fmeasure among the baselines, but poor accuracy. The baseline *Random* guesses true 50% of the time. We also constructed a baseline that guesses true/false over the test set based on the distribution in the training data (*Distribution*). This baseline is

smarter (more informed) than the other baselines, as it does not indiscriminately guess any one of the classes.

The bottom row of Table 20 shows the performance of our system (*System*). The skewness of the data affects the baselines as well as our system. Our system outperforms the best baseline Fmeasure by over 10 percentage points, and the best baseline precision by 14 percentage points. Comparing it to the baseline which has comparable accuracy, namely *Distribution*, we see that our system improves in Fmeasure by 24 percentage points.

The results indicate that by using simple features to capture target relations, it is possible to determine if two opinion sentences are related in the discourse better than baseline methods. In the next section, we will perform a more fine grained distinction. We will use the linguistic features explored here for recognizing target and opinion relations – that is, detecting their presence as well as determining their type.

6.2 RECOGNIZING DISCOURSE-LEVEL RELATIONS

Our preliminary studies showed that opinion frame relations can be detected above baseline. In this section we will recognize each of the discourse-level relations separately, and also classify their type.

For this, we will implement discourse-level relation recognizers using local as well as global discourse information. The idea here is that, while there are clues for recognizing discourse relations in the local context, in the global context, opinions and their polarities determine discourse-level relations and conversely, discourse relations help to disambiguate polarity.

To explain this in more detail, recall that, in our annotated corpus, discourse relations are established specific to opinions and their targets. A target, by definition, exists by virtue of the opinion that it is about. Target relations are the building blocks of opinion frame relations. The presence of opinions is thus a requirement for the existence of targets, target relations and opinion frame relations. Hence, the presence/absence of an opinion, and its polarity should be a useful indicator of whether a discourse relationship exists.

Furthermore, in Chapter 5, we found that polarity recognition is aided by the knowledge of discourse relations. The recognition of discourse-level relations and opinions are thus very interdependent. The interdependent nature of the relations is captured in the graphical structure in Figure 9. The DA instances form the nodes in this graph. The target relations and the discourse-level opinion relations (frame relations) form the links between the nodes. The information in the nodes (their opinion polarity) can be used to determine the links (whether there is a target or opinion relationship, and the link type), and the links help to disambiguate the values of the nodes. We build our global classification framework with respect to this general idea of iterative interdependent disambiguation, where the node labels as well as the structure of the graph are predicted in a joint manner.

In particular, our interdependent interpretation framework has three main units: an instance polarity classifier (IPC), a target-link classifier (TLC), and a frame-link classifier (FLC). IPC classifies each node (instance), which may be a sentence, utterance or any other text span, as *positive*, *negative* or *neutral*. TLC determines if a given node pair has related targets and whether they are linked by a *same* or *alternative* relation. FLC determines if a given node pair is related via frames, and whether it is a reinforcing or non-reinforcing link. Local clues available for each classifier help classification of the clear cases. Global discourse information augments the local information to aid in further disambiguation.

We use a collective a classification framework (a more evolved version of ICA from the previous chapter) to implement our interdependent interpretation framework. As seen previously, in the collective classification framework, there are two sets of features. The first are local features which can be generated for each instance or link, independent of the links in which they participate, or the instances they connect. The local features for IPC, TLC and FLC are described in 6.2.2. The second set of features, the relational features, reflect neighborhood information in the graph. The relational features for IPC, TLC and FLC are described in 6.2.3. First, we will look at the details of the ICA algorithm in Section 6.2.1 below.

6.2.1 Inter-dependent Interpretation Framework

We choose a collective classification algorithm implemented by our UMD collaborators Lise Getoor and Galileo Namata[1], for performing the inter-dependent interpretations on the graph in Figure 13. The collective classification algorithm used in this section is a a variant of the iterative classification algorithm (ICA) proposed by Bilgic et al [Bilgic et al., 2007]. It combines several common prediction tasks in graphs: node classification (predicting the label of an node) and link prediction (predicting the existence and class of a link between nodes). For our tasks, node classification directly corresponds to predicting opinion polarity and link prediction corresponds to predicting the existence of a *same* or *alternative* target link or a *reinforcing* or *non-reinforcing* frame link between nodes.

The pseudocode for the algorithm is shown in Figure 13. The ICA algorithm begins by predicting the node labels and the links using only the local features. The sets of all nodes and links are then randomly ordered and in turn, each node (or link) is predicted using the local features and the values of the currently predicted relational features based on previous predictions. This is repeated until some stopping criterion is met. For these experiments, the number of iterations are set to 30, which is sufficient for ICA to converge to a solution for most previous data sets.

The algorithm is one very simple way of making classifications that are interdependent. Once the local and relational features are defined, a variety of classifiers can be used. This algorithm is implemented using SVM classifiers from Weka [Hall et al., 2009].

6.2.2 Local Features

We use Dialog Act segmentation of the data, similar to the experimental set up in Chapter 5. Each of the classifiers, IPC, TLC and the FLC use local features and relational features. The local features for IPC are exactly the same as that for Local in Chapter 5. That is, the features are constructed using unigrams, sentiment and arguing lexicons, and DA tags. TLC and FLC both rely on discourse features. The features described in Table 19 capture the discourse context. We create these features for the DA instances to act as local features for

[1]This algorithm is included for completeness and should not be considered a contribution of the thesis.

for each opinion o **do** {bootstrapping}
  Compute polarity for o using local attributes
end for
for each target link t **do** {bootstrapping}
  Compute label for t using local attributes
end for
for each frame link f **do** {bootstrapping}
  Compute label for f using local attributes
end for
repeat {iterative classification}
  Generate ordering I over all nodes and links
  for each i in I **do**
    if i is an opinion instance **then**
      Compute polarity for i using local and relational attributes
    else if i is a target link **then**
      Compute class for i using local and relational attributes
    else if i is a frame link **then**
      Compute class for i using local and relational attributes
    end if
  end for
until Stopping criterion is met

Figure 13: DLOG-ICA Algorithm implemented by the UMD team

Feature	Task
Time difference between the node pair	TLC, FLC
Number of intervening instances	TLC, FLC
Content word overlap between the node pair	TLC,FLC
Focus space overlap between the node pair	TLC, FLC
Bigram overlap between the node pair *	TLC, FLC
Are both nodes from same speaker *	TLC, FLC
Bag of words for each node	TLC, FLC
Anaphoric indicator in the second node	TLC
Adjacency pair between the node pair	FLC
PDTB Discourse relation between node pair *	FLC

Table 21: Features and the classification task it is used for; TLC = target-link classification, FLC = Frame-link classification. The features indicated with a '*' were not present for our preliminary experiments in Section 6.1.2

TLC and FLC. Table 21 lists the features used by TLC and FLC. As seen in the table, we use all the features from the preliminary experiments for TLC and FLC in addition to some new features. Also, not all features are used for both classifiers – the *anaphoric indicator* feature is specific to target relations and is used exclusively for TLC; the *adjacency pair* feature is used only for FLC.

The features indicated with a '*' were not present for our preliminary experiments. The *bigram overlap* feature is introduced in an attempt to capture greater alignment between speakers. Our preliminary experiments were carried out on sentence pairs belonging to the same speaker. For the current experiments, we create DA pairs between instances belonging to the same as well as different speakers and we encode the speaker information as a feature (the *are-both-nodes-from-same-speaker* feature) into the link classifiers.

We add a new set of local features for FLC: the Penn Discourse TreeBank (PDTB) discourse relation between the pair. We have observed, in Chapter 3, that sometimes, there is an overlap between our opinion frame relations and the PDTB relations. The PDTB features we encode attempt to capture cases of contrast, elaboration etc. in the discourse. We group the list of discourse relations from PDTB into the following sets: Expansion, Contingency, Alternative, Temporal, Comparison. Each discourse relation in PDTB is associated with a list of discourse connective words. The PDTB provides a list of discourse connectives and the list of discourse relations that each connective signifies. Given a node pair, if the first word of the later instance (or the last word first instance) is a discourse connective word, then we assume that this node is connecting back (or forward) in the discourse and the feature set to which the connective belongs is set to true (e.g., if a latter instance is "because we should ...", it starts with the connective "because", and connects backwards via a Contingency relation).

6.2.3 Relational Features

The relational features used by IPC, TLC and FLC are listed in Table 22. Relational features incorporate the global discourse information, that is, the related class information and the discourse-relation information. Each row in Table 22 represents a set of features. Features are generated for all combinations of x, y and z for each row. For example, one of the

features in the first row is *Number-of-neighbors-with-polarity-type-positive-that-are-related-via-a reinforcing-frame-link*. Thus, each feature for the polarity classifier identifies neighbors for a given node via a specific relation (z or y) and factors in their polarity values.

Similarly, both link classifiers use polarity information of the node pair, and other link relations involving the nodes of the pair. Opinions are an important factor that determine the annotation of relations in our corpus. The *Polarity-of-the-DA-nodes* feature captures the opinion information relevant to the links (for both TLC and FLC). Similarly, the presence of a different type of link between the nodes is also a good indicator. For instance, if a target link is present between two nodes, it is likely that a frame link could also exist (though this is not always the case; if one of the nodes is neutral, the frame relation will not exist). The features *Presence-of-a-frame-link-z-between-the-nodes* and *Presence-of-a-target-link-ybetween-the-nodes* capture this information for TLC and FLC respectively. Similarly, when there is a lot of repetition or reiteration in the discourse, the nodes will be related to a number of other nodes. Thus, the presence of links with other nodes can be indicative that the given node pair also has a relation. This information is captured by the features *Number-of-other-frame-links-z-involving-the-given-DA-nodes* and *Number-of-other-target-links-y-involving-the -given-DA-nodes* .

6.2.4 Experiments

Our relational features incorporate three pieces of global information from the discourse: the target links, the opinion frame links and the polarity of the nodes linked via these relations. Depending on where this discourse information comes from, we explore three different experimental settings:

Perfect relational information: In this case, all the global relational information comes from an oracle. Thus, the relational information is very reliable.

Partial relational information: In this condition, the global information for the instances of the same type (for example, polarity information for IPC, target link information for TLC and the frame link information for FLC) comes from the previous iteration of the relational classifier. That is, this information is obtained automatically. The remaining

Feature
Opinion Polarity Classification
Number of neighbors with polarity type x linked via frame link z
Number of neighbors with polarity type x linked via target link y
Number of neighbors with polarity type x and same speaker linked via frame link z
Number of neighbors with polarity type x and same speaker linked via target link y
Target Link Classification
Polarity of the DA nodes
Number of other target links y involving the given DA nodes
Number of other target links y involving the given DA nodes and other same-speaker nodes
Presence of a frame link z between the nodes
Frame Link Classification
Polarity of the DA nodes
Number of other frame links z involving the given DA nodes
Number of other frame links z involving the given DA nodes and other same-speaker nodes
Presence of a target link y between the nodes

Table 22: Relational features: $x \in$ {non-neutral (i.e., positive or negative), positive, negative}, $y \in$ {same, alt}, $z \in$ {reinforcing, non-reinforcing}

relational information for each classifier is provided by an oracle.

No oracle information: In this condition, no oracle is used. All the relational information is obtained automatically.

We will explain these three settings using the discourse-relation graph in Figure 14. This graph has 4 DAs with positive (nodes 1 and 4) and negative (nodes 2 and 3) polarities, a *same* target relation (between nodes 2 and 3), *alternative* target relations (between nodes 1 and 2 and between nodes 1 and 4), reinforcing opinion relations (between nodes 1 and 2 and between nodes 2 and 3), and a non-reinforcing relation (between nodes 1 and 4).

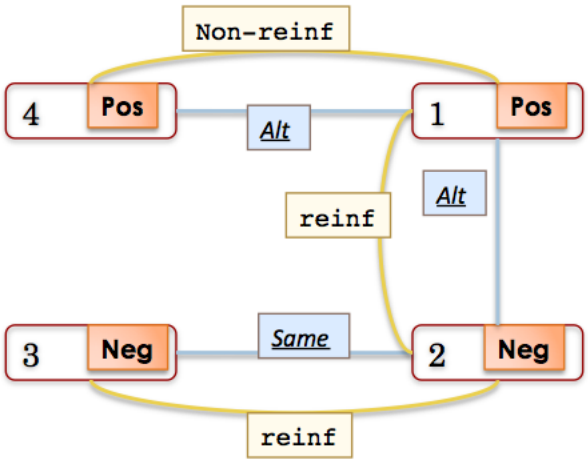

Figure 14: An example of a discourse-level opinion graph

Under the perfect relational information condition for IPC, the knowledge about the discourse relations and the polarity of the discourse neighbors is provided by an oracle (manual annotations). Figure 15 shows an example of prediction in this setting. As shown in Figure 15, if the IPC needs to predict the class of node 1, the discourse information that it is in an alternative and reinforcing relation with node 2, and that the node 2 has a negative polarity is provided by an oracle.

Figure 16 shows the perfect information condition for TLC. In this setting, the information of the nodes' polarity and all the discourse relations (except the one being predicted) is provided by an oracle. Specifically, as illustrated in Figure 16, if TLC is making a classification for the node pair node-1-node-2, every other piece of relational information in the graph

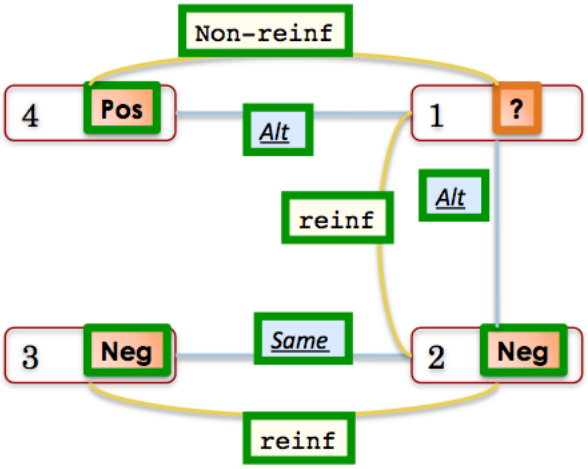

Figure 15: Perfect information condition for IPC. The information in green boxes is provided by the oracle.

is provided by the oracle. Similarly, Figure 17 illustrates the perfect information condition for FLC. Here, when FLC attempts to predict the discourse relation between nodes 1 and 2, every other piece of relational information is given by the oracle.

In the partial information condition for IPC (see Figure 18), IPC uses the oracle information about the target links and the frame links. However, for the polarity of its neighbors, it uses its own predictions from the previous round. That is, when IPC has to predict the value of node 1, it will use its predictions from the previous round for the polarity of nodes 2, 3 and 4.

In the partial information condition for TLC, polarity and frame relations are provided by the oracle. TLC uses its own prediction from the previous round to get the target relation information. Figure 19 illustrates this situation. Similarly, Figure 20 illustrates the partial information condition for FLC.

Finally, under the no oracle condition, all relational information is predicted. Here, the node values as well as the graph structure has to be predicted. The IPC gets the target relation information from TLC, frame relation information from FLC and polarity of the neighbors from its own predictions in the previous round. This setting is illustrated in

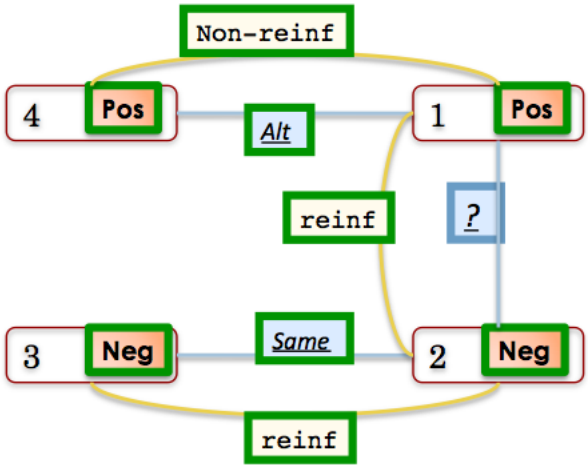

Figure 16: Perfect information condition for TLC. The information in green boxes is provided by the oracle.

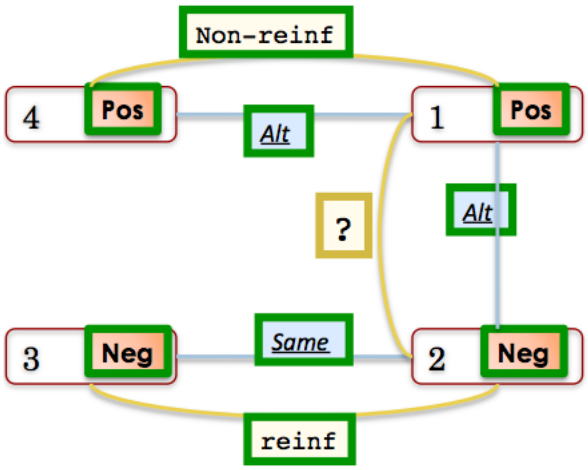

Figure 17: Perfect information condition for FLC. The information in green boxes is provided by the oracle.

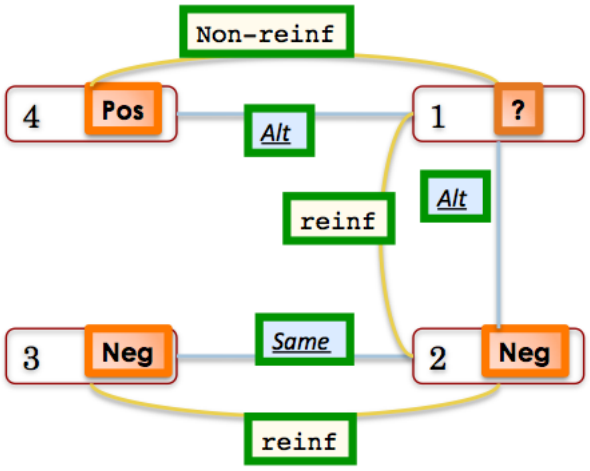

Figure 18: Partial information condition for IPC. The information in green boxes is provided by the oracle. The information in the orange boxes is predicted by IPC.

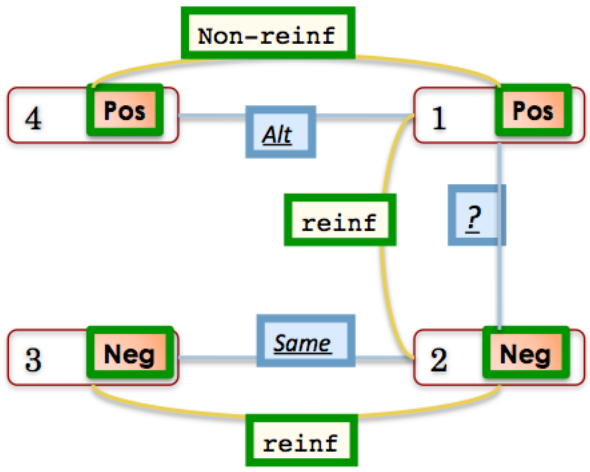

Figure 19: Partial information condition for TLC. The information in green boxes is provided by the oracle. Information in the blue boxes is predicted by TLC.

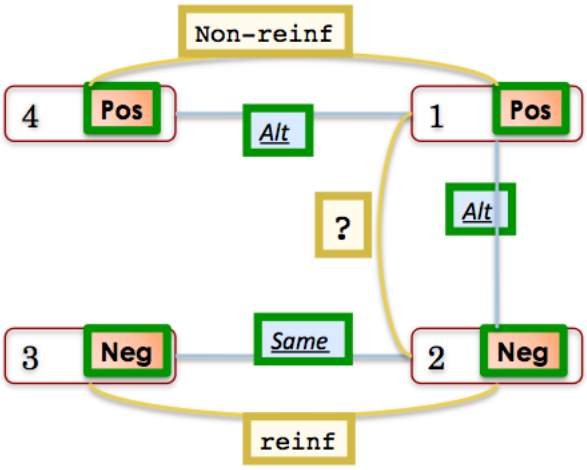

Figure 20: Partial information condition for FLC. The information in green boxes is provided by the oracle. Information in the yellow boxes is predicted by FLC.

Figure 21. Similarly, Figures 22 and 23 illustrate the no oracle condition for TLC and FLC.

6.2.5 Data

The data we use for experiments in this chapter is the same as that in Chapter 5. The annotated data consists of 7 scenario-based, multi-party meetings, and we use the DA segmentation as described in Section 5.1 (Chapter 5).

The link classifiers recognize relations between *DA pair* instances. We create DA pairs by first ordering the DAs by their start time, and then pairing a DA with five DAs before and after it. The classes for target-link classification are *no-link, same, alt.* The gold standard target-link class is decided for a DA pair based on the target link between the targets of the opinions contained in that pair. Similarly, the labels for the frame-link labeling task are *no-link, reinforcing, non-reinforcing.* The gold standard frame link class is decided for a DA pair based on the frame between opinions contained by that pair.

In our data, of the 4606 DAs, 1118 (24.27%) participate in target links with other DAs, and 1056 (22.9%) form opinion frame links. The gold standard data for links, which has

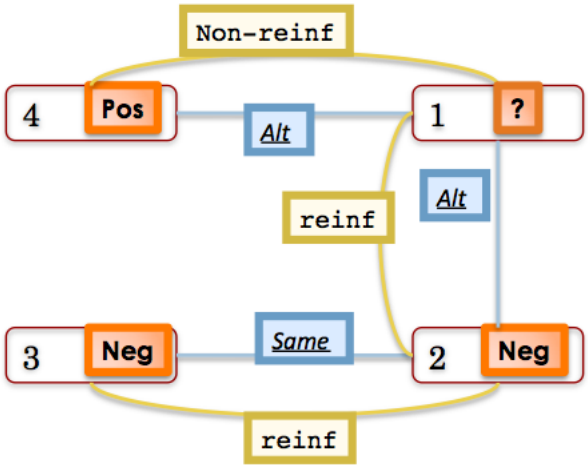

Figure 21: No oracle condition for IPC. The information in orange boxes is predicted by IPC, information in blue boxes is predicted by TLC and information in yellow boxes is predicted by FLC

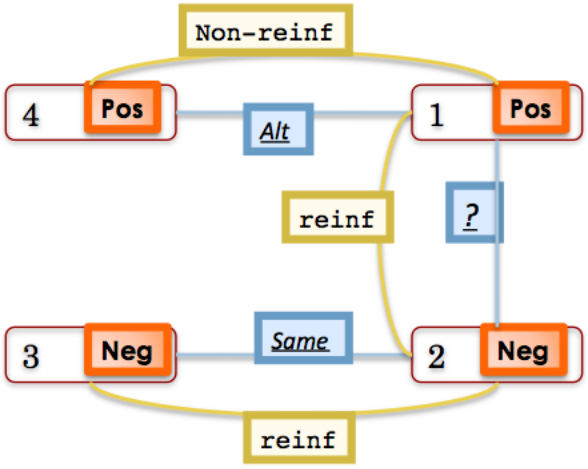

Figure 22: No oracle condition for TLC. The information in orange boxes is predicted by IPC, information in blue boxes is predicted by TLC and information in yellow boxes is predicted by FLC

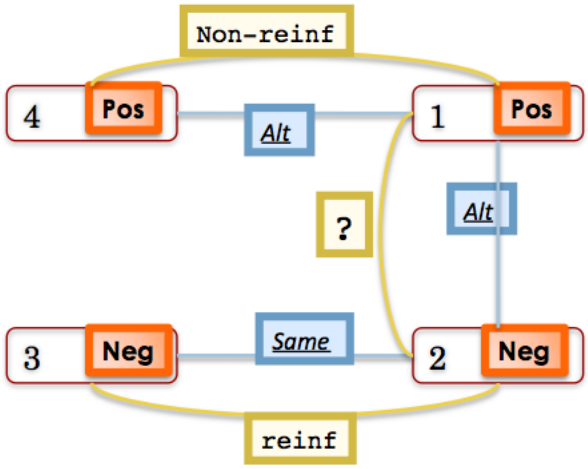

Figure 23: No oracle condition for FLC. The information in orange boxes is predicted by IPC, information in blue boxes is predicted by TLC and information in yellow boxes is predicted by FLC

pair-wise information, has a total of 22,925 DA pairs, of which 1371 (6%) pairs have target links and 1264 (5.5%) pairs have frame links.

We perform 7-fold cross-validation experiments using the 7 meetings. In each fold, 6 meetings are used for training and one meeting is used for testing. The results are averaged over the folds.

6.2.6 Classifiers

A simple baseline (Base) classifies the test data based on the distribution of the classes in the training data. Distribution-based baselines are implemented for all three classifiers.

For our local classifiers, we used classifiers from the Weka toolkit [Witten and Frank, 2002]. For opinion polarity, we used the Weka's SVM implementation. For the target link and frame link classes, the huge class skew caused SVM to learn a trivial model and always predict the majority class. To address this, we used a cost sensitive classifier in Weka where we set the cost of misclassifying a less frequent class, A,

to a more frequent class, B, as $|B|/|A|$ where $|class|$ is the size of the class in the training set. All other misclassification costs are set to 1.

For our collective classification, we use the above classifiers for local features (l) and use similar, separate classifiers for relational features (r). For example, we learned an SVM for predicting opinion polarity using only the local features and learned another SVM using only relational features. For the No Oracle condition, where we use IPC, TLC and FLC classifiers for generating the relational features, we combine the predictions using a weighted combination where $P(class|l,r) = \alpha * P(class|l) + (1-\alpha) * P(class|r)$. This allows us to vary the influence each feature set has to the overall prediction. The results for ICA No Oracle setting are reported on the best performing α (0.7).

6.2.7 Results

We use accuracy, precision, recall and Fmeasure metrics to measure the performance of the systems. As each classification task is a 3-way classification, we calculate the macro averages of precision, recall and Fmeasure. Macro average is calculated by first finding the value for each class and then averaging it over all the classes.

The classification results for target relations and opinion frame relations for the three conditions are reported in Table 23. For target link classification, we observe that the distribution-based baseline has good accuracy due to the class skew. However, Local TLC has significantly better precision ($p < 0.01$), recall ($p < 0.01$) and Fmeasure ($p < 0.05$) than this baseline . Similarly, for opinion frame relations, we see that the baseline has better accuracy than Local FLC. But Local FLC has better precision ($p < 0.01$), recall ($p < 0.01$) and Fmeasure ($p < 0.01$).

These results indicate that our linguistically motivated features are useful for recognizing target and opinion frame relations.

Moving on to our global classification settings (ICA in the Table 23), under the perfect information condition (Perfect Info), TLC as well as FLC have a substantial and significant improvement across all metrics over the corresponding distribution based baselines ($p < 0.001$). For the Partial Info condition too, TLC and FLC produce significant improvements

across all metrics ($p < 0.001$). Finally, when there is no oracle information (No Oracle column in Table 23), TLC and FLC perform similar to their local counterparts, and have similar (significant) improvements over the corresponding distribution-based baselines.

Now let us examine if the global classification methods produce improvements over the local methods. That is, whether the ICA-based methods for the link classifiers perform better than their local counterparts.

For the Perfect Information (Perfect Info) condition, TLC and FLC perform significantly better than the corresponding local classifiers. Perfect Info TLC improves substantially over Local TLC over all metrics ($p < 0.001$). Similarly, Perfect Info FLC also shows pronounced and significant improvements over Local FLC for all metrics ($p < 0.001$).

These results indicate that global relational information is useful over an above the local discourse information for both target link classification and frame link classification tasks.

Moving on to the Partial Info condition, we see a similar trend as the Perfect Info condition. TLC and FLC perform substantially better than their local counterparts across all metrics ($p < 0.001$). In fact, there is no difference in the performance between Partial Info TLC and Perfect Info TLC, or between Partial Info FLC and Perfect Info FLC. This indicates that, for the global target link classifier, perfect knowledge of the frame relations and the opinions is sufficient to produce the improvements. Similarly, the global frame link classifier achieves its improvements based on perfect knowledge of the target relations and opinions.

Both TLC and FLC under the No Oracle condition perform marginally better than their local counterparts across all metrics. However, this improvement is not statistically significant. This result indicates that, while the global classification framework can produce improvements over the local methods, it is essential that the relational information be reliable. When the relational information is noisy, the behavior of the global link classifiers is not significantly different from their local counterparts.

We will now turn to the measurement of polarity disambiguation under each of our experimental conditions. Table 24 reports the results. We have already performed a detailed comparison between the local and distributional methods for polarity classification in the previous chapter. Note that the experimental setting for global inference (ICA) in the

			ICA		
	Base	Local	Perfect Info	Partial Info	No Oracle Info
TLC					
Accuracy	88.5	85.8	98.1	98.2	86.3
Macro Precision	33.3	35.9	76.1	76.1	36.3
Macro Recall	33.3	38.1	78.1	78.1	38.1
Macro Fmeasure	33.1	36.0	74.6	74.6	36.5
FLC					
Accuracy	89.3	86.2	98.9	98.9	87.6
Macro Precision	33.3	36.9	81.3	82.8	38.0
Macro Recall	33.4	41.2	82.2	84.4	41.7
Macro Fmeasure	33.1	37.2	80.7	82.3	38.1

Table 23: Performance of Target Link Classifier (TLC) and Frame Link Classifier (FLC)

previous chapter, where there is perfect information about the links in the discourse but the polarities of the neighbors come from the predictions of the previous round, is the same as the Partial Info condition in Table 24.

We see that, under the Perfect Info condition the performance of the polarity classifier is the highest, across all metrics. This is expected as all the neighborhood information is reliable. However, under the No Oracle condition, the performance of IPC is not much different than the local classifier. This result indicates that, when the relational information is noisy, global polarity classification behaves similar to local polarity classification.

6.3 DISCUSSION

Our results for polarity classification using the ICA framework in the previous section indicate that our relational features are useful, but only when they are not noisy. Otherwise, the

			ICA		
	Base	Local	Perfect Info.	Partial Info	No Oracle Info
Accuracy	45.9	68.7	78.8	72.9	68.4
Macro Precision	33.3	61.6	73.1	65.9	62.1
Macro Recall	33.3	58.3	70.6	61.4	55.9
Macro Fmeasure	33.0	58.7	71.5	62.6	57.2

Table 24: Performance of IPC for the three conditions

performance achieved by ICA-based global methods is similar to that from local methods.

The polarity classifier (IPC) that uses automatically produced target and frame links does no better than the local classifier. This is because of the fact that the fully automated link classifiers have a large bias towards the "no link" class. There is a large skew in the target link data – 94% of the data has the "no link" label. Similarly 95% of the frame link data has the "no link" label. This causes each of the automated TLC and FLC to have a large bias towards guessing the "no link" class. Thus, the nodes are not able to see their discourse neighbors. Due to this, the performance is the same as that using local features.

In order to improve fine-grained polarity recognition systems using global information, we will thus need to improve the performance of target link classifiers and opinion frame link classifiers. While creating DA pairs for link classification, we pair each DA with all DAs within a window of size 5. The DAs in the corpus do have long distance relationships, but these are not prolific. In an attempt to capture long distance relations, we create a large number of pairs with "no link" classification, which skews the data. One way to circumvent the data skewness problem is to pair a DA with only one or two adjacent DAs.

Data skewness can also be addressed by designing a filtering strategy during data preparation. DA pairs that are clear cases of "no link" can be determined heuristically (for example, when one of the DAs does not contain content words) and can be filtered out to create a more balanced training data.

Another apparent problem, we believe, is strong coupling between classifiers in the cur-

rent setup. Noisy classifications from one classifier affect the bootstrapping process of other classifiers. For example, if the polarity classifier gets the discourse information of "no link" from the target-link and frame-link classifiers, it is not able to "see" its discourse neighbors. Thus, its discourse-based classification will be no different from its local classification. Similarly, a "no link" classification produced by the target-link classifier prevents the frame-link classifier from seeing its neighbors. As a result, there is no global influence that can change its classifications. One way to overcome this problem might be to limit the bidirectional inter-dependency to only the clear cases.

The performance of the local target link classifier can also be improved by providing *more comprehensive target annotations*. Recall that target annotations are created with respect to the opinions that they are about. That is, items are marked as targets only when they are targets of opinions. When annotated as targets, they can be candidates to be linked to similar such targets (or alternative targets). However, if an entity that has been previously annotated as target occurs again in the corpus without an associated opinion it will not be annotated. Consequently, no *same* target relation will exist between the two occurrences of the same entity. That is, targets need not be linked to other similar entities, even when they corefer. Thus, in this situation, two instances that share content words and corefer will have a "no link" label. So the target link learner will see different labels (*no-link*, *same*, *alternative*) for the same type of feature.

Note that this situation is not an inconsistency in our annotation scheme, as our scheme is not intended for annotating all forms of coreference. However, as the learner relies on features that are based on coreference, creating annotations that provide the learner with consistent coreference information will be useful.

This can be done by adding an additional annotation round after corpus annotations for discourse-level opinion relations are completed. In this round, all text spans that are *same* or *alternatives* to already annotated targets could be annotated as targets, and the corresponding links created. The additional annotations will not only give more consistent features to the learner, but will also reduce the class skew for the target link classifier.

6.4 RELATED WORK

In this chapter, we focus on detecting the discourse-level relationships such as target relations and opinion frame relations. In the field of product review mining, sentiments and features (aspects or targets) have been mined (for example, Yi et al. [Yi et al., 2003b], Popescu and Etzioni [Popescu and Etzioni, 2005], and Hu and Liu [Hu and Liu, 2006]). More recently there has been work on creating joint models of topic and sentiments [Mei et al., 2007, Titov and McDonald, 2008] to improve topic-sentiment summaries. We do not model topics; instead we directly model the relations between targets. The task of finding co-referent opinion topics by [Stoyanov and Cardie, 2008b] is similar to our target link classification task. However, our definition of targets is not limited to the topics; they may be propositions or events. Thus, our target-link classification looks at more general relationships.

Much of the other literature relevant to this work, such as global classification methods for sentiment analysis, has been discussed in the previous chapter.

6.5 SUMMARY

In this work, we explore if discourse-level relations can be automatically identified.

First, we explored the simplified problem of detecting the presence of opinion frame relations between opinion-bearing sentences. For this, we employed different types of linguistically motivated clues that captured discourse continuity, coreference and dialog relations. The resulting system is able to detect the presence of frame relations better than distribution-based baselines. Then, we explored the recognition (3-way classification) of target and opinion relations. For this recognition task we explored linguistically motivated local features and global relational features.

Our target relation recognizer employing discourse-motivated local features performs significantly better than the distribution based baseline. Similarly, our opinion frame relation classifier that uses linguistic features is able to perform better than the corresponding

distribution-based baseline. These results support the hypothesis 6.a, that our discourse-relation classifiers constructed using a variety of linguistic features perform better than baseline methods.

In the global classification framework, we augmented the existing features for each classifier (the target relation classifier and the frame relation classifier) with relational features. TLC and FLC classifiers under each of the experimental settings performed better than the distribution based baseline. When compared with their local counterparts, we found that global TLC and global FLC can produce improvements, when all or at least some of the relational information is reliable. However, if the relational information is noisy, as is the case when all of the relational information is generated automatically, the performance is no better than the local counterparts. Thus, our hypothesis 6.b, which states that discourse-relation classifiers constructed using global information perform better than local baselines is supported only when the global information is manually provided.

We also explored fine-grained polarity classification using the automatically generated frame and target relations. Here too, we found that when all of the relational data is generated automatically, the classifier is not able to achieve any improvement over the local polarity classifier. Overall, our results indicate that, while our local link classifiers perform better than distribution-based baselines, this performance is not sufficient to bootstrap the global link classifiers and produce an improvement in fine-grained polarity classification.

We discussed the factors that make fully automatic link classifications challenging. We talked about the different ways in which the classifier performance may be improved in the future. Specifically, we suggested changes in the way the data for link classifiers is created and changes in the classifier set up that may mitigate the currently identified drawbacks. We also suggest an addition to the annotated corpus to create more reliable features for learning. Future work in this area will explore these directions.

7.0 STANCE CLASSIFICATION IN PRODUCT DEBATES

In this chapter we will perform stance classification using opinion analysis. Stance refers to an overall position held by a person towards an object or proposition. For example, in a debate "iPhone versus Blackberry, which is better", a person may take a pro-iPhone or a pro-Blackberry stance. Similarly, being pro-choice, believing in evolution, supporting universal healthcare are all examples of stances. Stances are at a coarser level than our discourse-level relations. For instance in Example 1.2, the opinions collectively reveal a pro-rubbery material stance, and in Example 1.3 the opinions reveal a pro-curved shape (or anti-square shaped) stance.

In this work, we explore the hypothesis that *the ideas behind the discourse-level relations explored in this thesis are useful for stance classification in product debates.*

For our work on stance classification, we use online debates. Online debates are web forums where, for each debate, participants take sides and support their stances via justifications, reasons and opinions. The participants also self-report the side they support. This gives us annotated stance data for our experiments. Using a different data set also gives us the opportunity to test whether our ideas of discourse-level relations are applicable to a new data set. Testing on user-reported stances allows us to get an out-of-lab, real world validation for our ideas. As compared to speech data, web texts have fairly good writing, which enables us to create parse trees and explore syntactic patterns for finding targets of opinions. Moving to the web also enables us to employ web mining for learning relations.

Thus, under the main high-level goal of stance recognition, we will also pursue a number of additional goals in this work:

Removing reliance on human annotation: Up till now, our work on discourse-level relations has involved human effort. We performed human reliability studies in Chapter 4,

utilized the human-generated discourse annotations for polarity analysis in Chapter 5, and used the human generated corpora for supervised machine learning experiments in Chapter 6. In this chapter, we would like to remove all human annotation efforts and try to solve the problem of stance classification in an unsupervised fashion. Thus, one of our additional goals is to learn the elements of our discourse-level relations without employing manual annotations.

Studying discourse-level relations in real world data: Online debate data is very different from the scenario-based AMI meetings that have been our test bed up till now. By mining weblogs and forums, and testing on debate forums, we have an opportunity to test if and how our discourse-level relations are manifested in real-world scenarios.

Exploring more domains: Previously, the scenario-based meetings limited our domain to that of TV remote controls. Now, as we use web forums, we can explore more domains such as browsers, phones, operating systems, video games and other products. That is, we aim to preform stance classification on different product domains.

Finding targets of opinions: Our experiments up till now did not attempt to explicitly find the targets of opinions, even though target annotations are a part of our linguistic scheme. This is because multi-party face to face conversations are rife with disfluencies, restarts, interruptions and ungrammatical constructs that make parsing very unreliable. As a result, it is not possible to use syntactic dependencies to find the targets (for example, finding subject of an opinion word or direct object of an evaluative term). Relative to this, the web data is more tractable. As a result, we will be able to explore ways to find targets of opinions.

Finding target relations by unsupervised method: In the previous chapter, we attempted to learn target relations using a supervised approach. As the web data has no manual annotations corresponding to our target relations, we explore if it is possible to capture these relations (or a subset of them) using web mining.

Explicitly finding reinforcing/non-reinforcing relations: In the previous chapters, discourse-level reinforcing and non-reinforcing relations exist based on the opinion frames that are constructed using target relations. Thus, target relations were necessary for finding reinforcement/non-reinforcement between opinions. We have seen that *same* target

relations overlap with a number of linguistic phenomena such as coreference, synonymy, is-a relationships and many forms of bridging descriptions. Thus, in trying to find target relations, we are also attempting to find all these types of linguistic relations. Now, the question is do we really need to explicitly solve all these linguistic problems before we can find discourse-level opinion relations, or is there a way to capture the *same/alternative* information more implicitly? Thus, in this work, we attempt to learn the reinforcing/non-reinforcing relations independent of explicit target relations.

Thus, under our high-level hypothesis, we have two specific hypotheses:

7.a The system that automatically learns and employs the elements of the target relations performs better than baseline for stance classification.

7.b The system that automatically learns and employs the elements of the discourse-level opinion relations performs better than baseline for stance classification.

Notice that we learn elements of the target and opinion relations, that is, we do not learn *all* the relations defined in our linguistic scheme in Chapter 3. This is primarily due to the fact that our unsupervised approaches are more amenable to learning certain types of relations (for example, web mining can be used to find relatedness between words, but using this to find alternatives in a discourse is difficult). Also, as our test bed is stance classification of product debates, we only need to pay attention to the relations that create an impact on stance classification in our data.

The work in this chapter is markedly different from the previous chapters. Primarily, we see a difference in the task (stance classification), in the data (online debates), and in the approach (unsupervised systems).

The rest of this chapter is organized as follows. In Section 7.1 we will first observe how people express stances in online debates and develop an intuition on why the discourse-level relations in this thesis might be useful for stance classification. We will then describe the basic units, the *opinion target pairs*, essential for all our systems in Section 7.2. We describe our unsupervised system that finds relations between targets in Section 7.3. The system that employs web mining to learn the discourse-level opinion relations is described

in Section 7.4. Section 7.5 describes the system that employs these relations to perform debate stance classification. Section 7.6 presents our experiments and Section 7.7 presents qualitative analysis and discussion. We discuss related work in Section 7.8 before concluding in Section 7.9.

7.1 STANCES IN ONLINE PRODUCT DEBATES

Online debates are goal-oriented discussions about hot topics, where participants express their support for one side or the other. In this chapter we analyze stances in online dual-sided, dual-topic debates about products. The debate topic sets up a choice between two products, and each side corresponds to supporting one of the products. An example of such a debate is "iPhone vs. Blackberry," where $topic_1$ = iPhone, $topic_2$ =Blackberry, $side_1$ = pro-iPhone, and $side_2$=pro-Blackberry. Let us look at a snippet from a post from this debate on http://www.convinceme.net. This post reveals a pro-iPhone stance.

(7.1) *iPhone* **of course**. *Blackberry* is now for the **senior businessmen market**! The *iPhone* **incarnate the 21st century** whereas *Blackberry* symbolizes an **outdated technology**. The *iPhone* can **reach a very diversified clientele**

Here the participant, in supporting the pro-iPhone stance, justifies why the side he supports is a better option. Accordingly, there are multiple positive opinions towards the iPhone. Another key strategy in debates is to argue why the opposing side is not good. We see this in the above example – the participant uses multiple negative opinions towards the opposing choice, the Blackberry, to further reinforce his pro-iPhone stance. The variety of opinions is used to reinforce an overall pro-iPhone stance. Notice that this snippet is not unlike the snippet from the AMI meeting in Example 1.3 where the speaker uses a variety of opinions to reinforce an overall pro-curved shape stance.

While the online debate genre is very different from face-to face multiparty conversations, there are some similarities between AMI meetings and online debates. Similar to AMI meetings, online debates are very goal oriented and hence the posts are focussed on a central

topic. Also, in both data sets, participants wish to support their stance and convince others about their choice; consequently the expressed opinions reveal their stance.

Given below are examples of debate posts. Example 7.2 is taken from the iPhone vs. Blackberry debate, Example 7.3 is from a Firefox vs. Internet Explorer debate, and Example 7.4 is from a Windows vs. Mac debate:

(7.2) While the iPhone may appeal to younger generations and the BB to older, there is no way it is geared towards a less rich population. In fact it's exactly the opposite. It's a gimmick. The initial purchase may be half the price, but when all is said and done you pay at least $200 more for the 3g.

(7.3) In-line spell check...helps me with big words like onomatopoeia

(7.4) Apples are nice computers with an exceptional interface. Vista will close the gap on the interface some but Apple still has the prettiest, most pleasing interface and most likely will for the next several years.

Online debates have a number of characteristics and challenges that highlight the importance of the elements of our discourse-level relations. We discuss each of these in the following sections.

7.1.1 Targets

In debates, what the opinions are about, their targets, become vital to determining the side. Debate participants, in advocating their choice, switch back and forth between their opinions towards the sides. Examples 7.1, 7.2 and 7.4 illustrate this phenomenon. In these examples, there are positive as well as negative opinions. Thus, the opinions by themselves are not informative. It is only when we know what the opinion is about, that we can infer the stance.

The importance of finding targets of opinions is accentuated due to the dual-topic, dual-sided nature of our data. For instance, in Example 7.1, there are positive as well as negative opinions. The positive opinions are directed towards the side (topic) that is being supported and the negative opinions are directed towards the side that is being opposed.

7.1.2 Target Relations

In debates, how the individual targets relate to the main topics becomes important in determining the stance. Examples 7.5, 7.6, 7.7 and 7.8 (from the iPhone vs. Blackberry debate) illustrate this. The first two examples are from pro-Blackberry posts and the remaining two are from pro-iPhone posts. Notice that none of the targets in these examples are either iPhone or Blackberry. Instead the participants choose to express their opinions about Pearl (a type of Blackberry), QWERTY keyboard (a feature of Blackberry), Apple (the maker of iPhones), and MAC OS (a feature of iPhone). It is only when we relate each of these targets to the main topics, iPhone and Blackberry, that we can determine the stance.

(7.5) The *Pearl* does music and video **nicely**

(7.6) First, you still **can't beat** the *full QWERTY keyboard* for **quick, effortless** *typing*.

(7.7) Well, *Apple* has always been **a well known company**.

(7.8) Its *MAC OS* is also a **unique thing**.

The above examples illustrate that targets mentioned in a post can have a *same* target relation with one of the main topics.[1] Thus, finding these relations can help with debate stance classification.

Furthermore, the debate title sets up the two topics as alternatives to each other for the discourse of the debate. The two topics are mutually exclusive options in the context of debate as participants can choose only one stance (at a time). In Examples 7.1, 7.2, and 7.4, the writers employ alternatives effectively – they argue for the topic they support and against the alternative topic to reinforce their overall stance.

7.1.3 Reinforcing Relations

Each of the debate stances are associated with some clear cases of opinions. For example, in an iPhone vs. Blackberry debate, the pro-iPhone stance is associated with positive opinions towards iPhones and negative opinions towards Blackberries. Similarly, a pro-Blackberry

[1] There may be target relations between non-topical targets within the post, but we do not explore these in this work.

stance is associated with positive opinions towards Blackberries and negative opinions towards iPhones. These opinions clearly map to the corresponding stances.

However, debate participants do not always mention the main topics explicitly, as we saw in Examples 7.5, 7.6, 7.7 and 7.8. In the previous section, we discussed finding target relations to remedy this problem. There is another approach that can handle these cases: finding reinforcing relations. That is, determining how the individual opinions in these examples relate to the clear cases of opinions representing the stances. For example, positive opinions towards Pearl *reinforce* positive opinions towards Blackberry; and positive opinions towards Apple reinforce positive opinions towards iPhone. That is, by finding if the opinion expressions reinforce the clear cases (positive/negative opinions towards iPhone/Blackberry), we can find the stances.

In Examples 7.5, 7.6, 7.7 and 7.8 each of the targets are *unique* to one side or the other. For example, in an iPhone vs. Blackberry discourse, Pearl and QWERTY uniquely identify Blackberry; and Apple and Mac are exclusively associated with iPhone. Thus, positive opinions towards Pearl and QWERTY automatically reinforce positive opinions towards Blackberry, and positive opinions towards Apple or Mac OS reinforce positive opinions towards iPhone. When there are such unique mappings, employing target relations might be sufficient to infer what stances the opinions support. In fact, a system employing the reinforcing relations discussed above and a system employing target relations discussed in the previous section might perform similarly.

However, there are scenarios where target relations are not sufficient to determine the stance. Let us consider the following example to illustrate this:

(7.9) **Faster** *keyboard* input.

As iPhones and Blackberries, both have keyboards, the target *keyboard*, is associated with both topics – it is a *shared aspect*. Thus, we need to explicitly determine if positive opinions towards keyboards are used to reinforce positive or negative opinions towards iPhones or Blackberries.

Shared aspects are not uncommon in debates – debates are generally regarding topics that belong to the same domain (e.g. cell phones, browsers, video games). Due to this, they

have many common or shared aspects. For example, both iPhones and Blackberries have e-mailing capabilities, both can be used to take photos, both have batteries and so on. Thus, when an opinion is expressed towards a shared aspect, we have to determine specifically whether this opinion is used to reinforce a pro-iPhone stance or a pro-Blackberry stance.

In general we found that, in spite of the increased complexity, shared aspects may be used to determine the debate stance. This is possible because of the following:

- Certain shared aspects may be generally perceived to be better on one side than the other. For example, e-mails are commonly perceived to be better on Blackberry (This was the case when the data was collected. This public perception may have since changed). Thus, bringing up good emailing capabilities is a way to strengthen a pro-Blackberry stance.
- Value for shared aspects depends on personal preferences. For example, music (or map), a common capability of mobile smart-phones, is typically valued more by people in the iPhone community. Thus, if a person is making a pitch based on great music capabilities, he is more likely to belong to the pro-iPhone side.

Thus, when we encounter an opinion towards a shared aspect, we need to find out how likely it is that this opinion is used to reinforce either one of the debate stances.

Notice that the opinion relations that we talk about are particularized to the targets (as we noted in Section 7.1.1, the opinions themselves do not convey stance information).

7.1.4 Concessions

In debates, the goal of a participant's post is to fortify his/her stance. Hence, it is safe to assume that most of the opinions expressed in a post are reinforcing an overall stance. However, a particular type of non-reinforcement, *concessions*, are common in debates. This is because while debating, participants often refer to and acknowledge the viewpoints of the opposing side. Examples 7.10 and 7.11 illustrate concessions.

(7.10) While the *iPhone* **looks nice** and **does play a decent amount of music**, *it* **can't compare** in functionality to the BB.

(7.11) I **like** my *music*, and *phone*, but I **don't want** to *carry a brick around in my pocket* when I only need my phone.

In Example 7.10, there are two positive opinions and one negative opinion towards the iPhone. This is a non-reinforcing discourse scenario as there are positive and negative opinions towards the same item. In fact, this is a special type non-reinforcement, where one opinion is conceded and the other is endorsed. The conceded opinions do not support the stance, in fact they are indicative of the stance that the writer *does not* support.

Concessionary scenarios are common in our product debates. Examples 7.2 and 7.4 also illustrate concessionary opinions.

7.2 OPINION TARGET PAIRS

In Section 7.1.1 we observed that opinions by themselves are not informative of the stance – their targets are vital elements. Furthermore, target mentions by themselves are also not enough for stance classification – a mention of an iPhone is not indicative of a stance, however a positive (or negative) opinion towards it is. It is the combination of the opinions and their targets that convey stance information. We thus pair opinions with their targets to create *opinion target pairs*, which are represented as $target^{polarity}$. Specifically, an opinion target pair consists of a target and the polarity of the associated opinion. We mask the actual opinion expression with its polarity in order to create a more general pattern. For instance, in Example 7.5, the opinion target pair is $Pearl^{+}$. In Example 7.1, there are three instances of the opinion target pair $iPhone^{+}$ and two instances of the opinion target pair $Blackberry^{-}$.

An additional motivation for creating opinion target pairs is that by encoding the target information along with opinions, we can attempt to find reinforcing and non-reinforcing relations between these units directly. For instance, in Example 7.5, a positive opinion towards Pearl ($Pearl^{+}$) is used to reinforce a positive opinion towards Blackberry ($Blackberry^{+}$). Similarly, $full\ QWERTY\ keyboard^{+}$ is used to reinforce $Blackberry^{+}$ in Example 7.6, $Apple^{+}$ is used to reinforce $iPhone^{+}$ in Example 7.7, and $MAC\ OS^{+}$ is used to reinforce $iPhone^{+}$ in Example 7.8. By constructing opinion target pairs, we do not need to explicitly find *same/alternative* relations between the targets, that is, we can *bypass* the target

relations.

Finally, opinion target pairs encode the information that is required for handling opinions towards shared aspects. We saw in Example 7.9 that, the keyboard is an aspect common to both iPhone and Blackberry, and thus shares a *same* target relation with both of them. In this situation, the target relation is not useful for determining what stance the positive opinion towards keyboard actually supports. Thus, we need to explicitly find how likely it is that a positive opinion towards keyboard ($keyboard^{+}$) reinforces positive opinions towards iPhone ($iPhone^{+}$) or positive opinions towards Blackberry ($Blackberry^{+}$).

7.2.1 Opinion Target Pair Construction

We need to find opinions and pair them with targets, both to mine the web for general preferences and to classify the stance of a debate post. To find opinions, we look up words in a subjectivity lexicon [Wilson et al., 2005a] [2]. This lexicon contains approximately 8000 words which may be used to express opinions. All instances of those words are treated as opinions. Each lexicon entry consists of a subjective word, its prior polarity (positive ($^{+}$), negative ($^{-}$), neutral (*)), morphological information, and part of speech information. An opinion is assigned the prior polarity that is listed for that word in the lexicon, except that, if the prior polarity is positive or negative, and the instance is modified by a negation word (e.g., "not"), then the polarity of that instance is reversed.

To pair opinions with targets, we built a rule-based system based on dependency parse information. The dependency parses are obtained using the Stanford parser.[3] We developed our syntactic rules on separate data that is not used elsewhere in this work. Table 25 illustrates some of these rules. Note that the rules are constructed (and explained in Table 25) with respect to the grammatical relation notations of the Stanford parser. As illustrated in the table, it is possible for an opinion to have more than one target. In such cases, a single opinion results in multiple opinion target pairs, one for each target. We create a total of 14 rules for opinion target pairing. Specifically, there are six rules for opinion words that are adjectives, four rules for opinions that are nouns, two for verbs and two for adverbs.

[2] Available at http://www.cs.pitt.edu/mpqa.
[3] http://nlp.stanford.edu/software/lex-parser.shtml.

DIRECT OBJECT Rule: dobj(opinion, target)
In words: The target is the direct object of the opinion
Example: I love$_{opinion1}$ Firefox$_{target1}$ and defended$_{opinion2}$ it$_{target2}$
NOMINAL SUBJECT Rule: nsubj(opinion, target)
In words: The target is the subject of the opinion
Example: IE$_{target}$ breaks$_{opinion}$ with everything.
ADJECTIVAL MODIFIER Rule: amod(target, opinion)
In words: The opinion is an adjectival modifier of the target
Example: The annoying$_{opinion}$ popup$_{target}$
PREPOSITIONAL OBJECT Rule: if prep(target1,IN) $\Rightarrow$ pobj(IN, target2)
In words: The prepositional object of a known target is also a target of the same opinion
Example: The annoying$_{opinion}$ popup$_{target1}$ in IE$_{target2}$ ("popup" and "IE" are targets of "annoying")
RECURSIVE MODIFIERS Rule: if conj(adj2, opinion$_{adj1}$) $\Rightarrow$ amod(target, adj2)
In words: If the opinion is an adjective (adj1) and it is conjoined with another adjective (adj2),
then the opinion is tied to what adj2 modifies
Example: It is a powerful$_{opinion(adj1)}$ and easy$_{opinion(adj2)}$ application$_{target}$
("powerful" is attached to the target "application" via the adjective "easy")

Table 25: Examples of syntactic rules for finding targets of opinions

Once these opinion target pairs are created, we mask the identity of the opinion word, replacing the word with its polarity. Thus, the opinion target pair is converted to a polarity target pair. For instance, "pleasing-interface" is converted to $interface^{+}$.

7.3 STANCE CLASSIFICATION EMPLOYING TARGET RELATIONS

As noted in Section 7.1.2 (Examples 7.5, 7.6 7.7 and 7.8), target relations are important. In this section, we will explore ways to find target relations by using Pointwise Mutual Information (PMI), and then employ these relations for stance classification.

There are two motivations for building a stance classifier based on target relations. First, we want to investigate if elements of our target relations can be learnt in an unsupervised fashion for product debates. Second, previous work in product review mining has employed relations between products and their aspects. We want to test if this approach is sufficient for debate stance classification and whether our discourse-based reinforcing relations can produce further improvements.

Figure 24 shows the schematic of our system. We parse the posts and use the process described in Section 7.2 to extract opinion target pairs for each opinion expressed in the post. Specifically, we use the opinion target pairing process to find all the targets in a debate post. For example, as shown in the figure, $email^{+}$ will be extracted from the sentence "I like email". We then find semantic relatedness of each target in the post with the two main topics of the debate by calculating Pointwise Mutual Information (PMI) between the target term and each topic over the entire web corpus. We limit the PMI search to nouns because we found, in our initial experiments on development data, that topics correspond to nouns, and PMI relatedness between the topics and words belonging to other parts of speech is noisy. In product debates, almost all the targets are nouns or noun phrases. Furthermore, all our rules for opinion target pairing result in nouns being extracted as targets.

In order to find semantic relatedness, we use the API provided by the Measures of Semantic Relatedness (MSR) engine hosted at http://cwl-projects.cogsci.rpi.edu/msr/. The MSR engine issues Google queries to retrieve documents and finds the PMI between any two

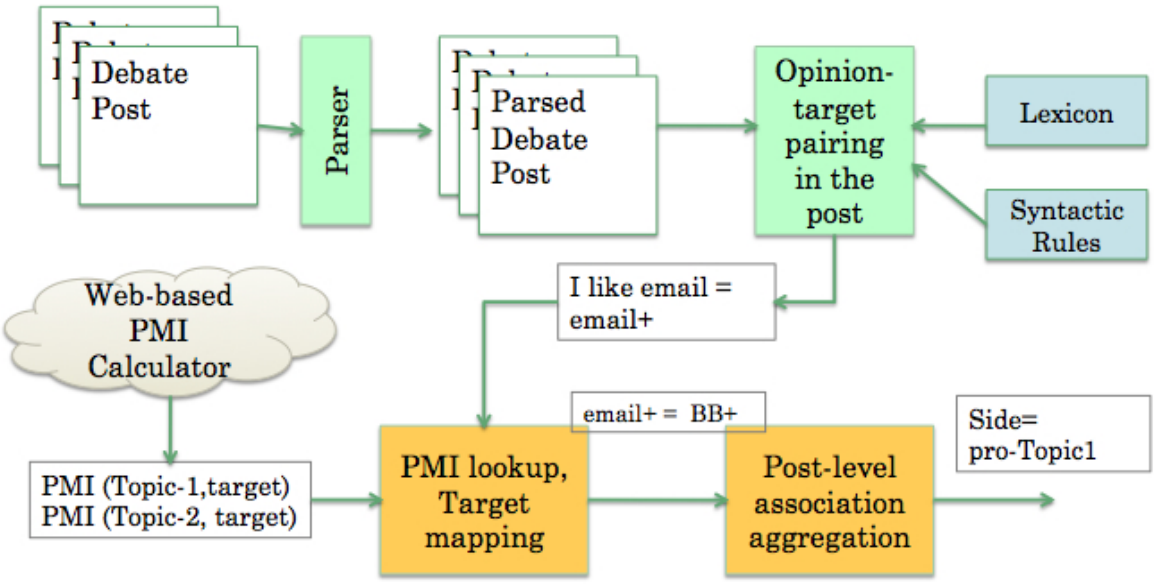

Figure 24: Schematic of the system learning target relations from the web for stance classification

given words. Table 26 lists PMIs between the topics (iPhone and Blackberry) and targets found in the debate posts.

Each target k is assigned to the topic with the higher PMI score. That is, it is considered to be a proxy for the topic it is closest to. Specifically, if

$$PMI(topic_1, k) > PMI(topic_2, k) \Rightarrow k = topic_1$$

and if

$$PMI(topic_2, k) > PMI(topic_1, k) \Rightarrow k = topic_2$$

For instance, in Figure 24, email is found to be closer to Blackberry; hence $email^+$ is replaced with $Blackberry^+$.

By using this approach, the system is able to learn a subset of the *same* target relations defined in Chapter 3. It is also able to handle cases of shared aspects. The PMI closeness value and the equations above will sever the *same* relation between the shared aspect and one of the topics. Thus, the shared aspect will be considered as *same* as only one of them.

Finally, we perform post-level aggregation and calculate the side score using Equations 7.12 and 7.13. Each post is assigned the side with the higher score.

$$score(side_1) = \#topic_1^+ + \#topic_2^- \tag{7.12}$$

$$score(side_2) = \#topic_1^- + \#topic_2^+ \tag{7.13}$$

7.4 LEARNING OPINION RELATIONS

In this section, we attempt to find reinforcing relations that are useful for recognizing stances in an unsupervised fashion. Our unsupervised method relies on web mining which is done using seed words extracted from the debate title. Opinions towards items in the debate topic are the clearest ways to argue for a stance. For example, in the iPhone vs. Blackberry debate, an obvious way to argue for pro-iPhone stance is to say positive things about iPhone

word	iPhone	blackberry
storm	0.923	0.941
phone	0.908	0.885
e-mail	0.522	0.623
ipod	0.909	0.976
battery	0.974	0.927
network	0.658	0.961
keyboard	0.961	0.983

Table 26: PMI of words with the topics

($iPhone^+$) and an obvious way to support a pro-Blackberry stance is to express positive opinions towards Blackberry ($Blackberry^+$). We can apply the idea of alternative relations to the debate topic to create additional seeds. The debate scenario sets up the two main topics as alternatives, and thus one may argue for one topic by arguing against the alternative. For example, in the iPhone vs. Blackberry debate, negative opinions towards iPhone ($iPhone^-$) reveal a pro-Blackberry stance and negative opinions towards Blackberry ($Blackberry^-$) reveal a pro-iPhone stance. Thus, the debate topic "iPhone vs. Blackberry" provides four opinion target pairs as seeds. Two of these map to a pro-iPhone stance ($iPhone^+$ and $Blackberry^-$) and the remaining two ($iPhone^-$ and $Blackberry^+$) map to pro-Blackberry stance. In general, in a dual topic debate regarding two topics $Topic1$ and $Topic2$, the opinion target pairs $Topic1^+$ and $Topic2^-$ correspond to a pro-Topic1 stance, and the opinion target pairs $Topic1^-$ and $Topic2^+$ correspond to a pro-Topic2 stance. These four opinion target pairs are called *opinion topic* pairs as these contain explicit topic mentions.

Debate posts contain a number of opinion target pairs, not all of which are explicit topic mentions. We use web mining to find associations between opinion-targets found in the posts and each of the four opinion-topics. These associations indicate how likely it is that an opinion-target in the post is used to reinforce the opinion-topics.

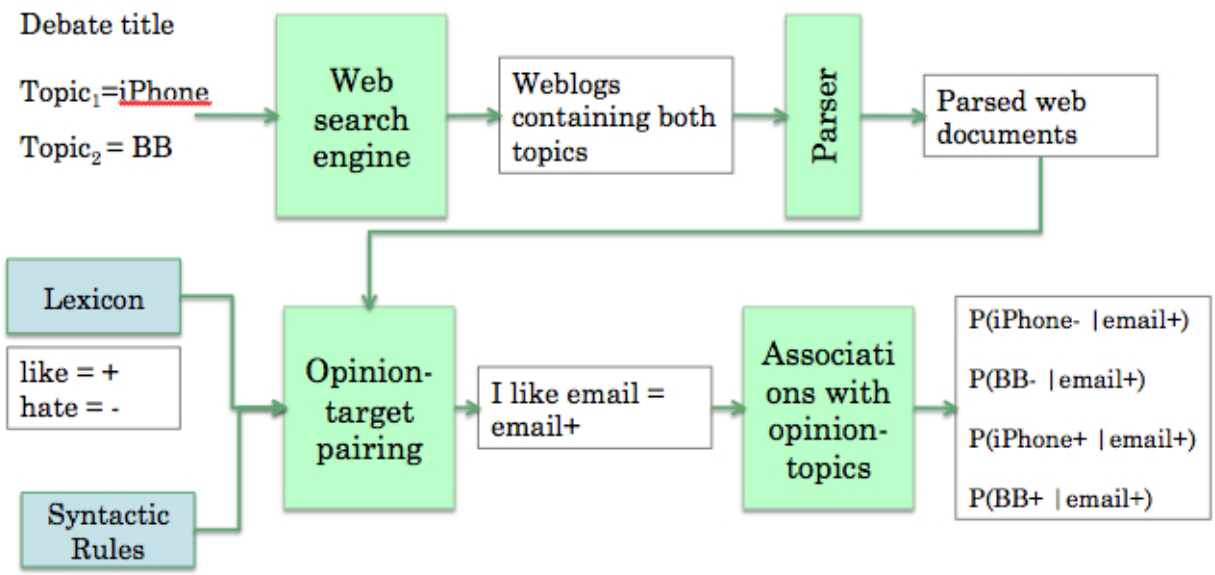

Figure 25: Learning relations from the web

The architecture of our system that learns associations between opinion-topics and opinion-targets is shown in Figure 25. For each debate, we start with the opinion topic pairs created from the debate title. We download weblogs and forums that talk about *both* the topics of that debate. For example, for the iPhone vs. Blackberry debate, we search the web for pages containing "iPhone" and "Blackberry." We used Yahoo's search API and imposed a search restriction that the pages should contain both topics in the http URL. This ensured that we downloaded relevant pages.

We parse each of these web documents using the Stanford parser and apply the method described in Section 7.2 for creating opinion target pairs. In this step, we find all instances of words in the lexicon, extract their targets, and mask the opinion words with their polarities. For example, suppose the sentence *"I like email"* is in the corpus. The system extracts the pair "like-e-mail", which is masked to "positive-email," which we notate as $email^{+}$.

Next, we find associations between opinion-targets and opinion-topics. If the target in an opinion target pair happens to be one of the topics, we select opinion target pairs in its

vicinity for further processing (the rest are discarded). The intuition behind this is that, if someone expresses an opinion about a topic, he or she is likely to follow it up (reinforce it) with reasons for that opinion. Sentiments in the surrounding context thus reveal factors that influence a preference or dislike towards the topic. We define vicinity as the same sentence plus the following 5 sentences, where the sentences are in a reinforcing discourse. We use the PDTB list of discourse connectives belonging to the Contrast, Concession and Contra-expectation categories to detect a non-reinforcing discourse. If a sentence in the vicinity has a discourse marker from this list, we exclude the opinion-targets found in it and all the following sentences from further processing.

Each unique target word $target_i$ in the web corpus, i.e., each word used as the target of an opinion one or more times, is processed to generate the following conditional probabilities.

$$P(topic_j^q | target_i^p) = \frac{\#(topic_j^q, target_i^p)}{\#target_i^p} \tag{7.14}$$

where $p = \{^+, ^-, ^=\}$ and $q = \{^+, ^-, ^=\}$ denote the polarities of the target and the topic, respectively; $j = \{1, 2\}$; and $i = \{1...M\}$, where M is the number of unique targets in the corpus. For example, $P(Mac^+|interface^+)$ is the probability that "interface" is the target of a positive opinion that is in the vicinity of a positive opinion toward "Mac."

Table 27 lists some of the probabilities learned by this approach. (Note that the neutral cases are not shown.)

7.4.1 Interpreting the Learned Probabilities

The probabilities in Table 27 align with what we qualitatively found in our development data. For example, the opinions towards Storm essentially follow the opinions towards Blackberry; that is, positive opinions toward Storm reinforce positive opinions toward Blackberry, and negative opinions toward Storm are usually found in the vicinity of negative opinions toward Blackberry (for example, in the row for $storm^+$, $P(blackberry^+|storm^+)$ is much higher than the other probabilities). Thus, an opinion expressed about Storm is usually the opinion one has toward Blackberries. This is expected, as Storm is a type of Blackberry. A similar example is $ipod^+$, which follows the opinion toward iPhone. This is interesting because an

	$side_1$ (pro-iPhone)		$side_2$ (pro-Blackberry)	
$term^p$	A	B	C	D
$storm^+$	0.227	0.068	0.022	0.613
$storm^-$	0.062	0.843	0.06	0.03
$phone^+$	0.333	0.176	0.137	0.313
$e\text{-}mail^+$	0	0.333	0.166	0.5
$ipod^+$	0.5	0	0.33	0
$battery^-$	0	0	0.666	0.333
$network^-$	0.333	0	0.666	0
$keyboard^+$	0.09	0.12	0	0.718
$keyboard^-$	0.25	0.25	0.125	0.375

Table 27: Probabilities learned from the web corpus (iPhone vs. Blackberry debate) A = $P(iPhone^+|term^p)$; B = $P(Blackberry^-|term^p)$; C = $P(iPhone^-|term^p)$; D= $P(Blackberry^+|term^p)$

iPod is not a phone; the association is due to preference for the brand. In contrast, the probability distribution for phone does not show a preference for any one side, even though both iPhone and Blackberry are phones. This indicates that opinions towards phones in general will not be able to distinguish between debate sides.

Another interesting case is illustrated by the probabilities for "e-mail". People who like e-mail capability are more likely to praise the Blackberry, or even criticize the iPhone — they would thus belong to the pro-Blackberry camp.

While we noted earlier that positive evaluations of keyboards are associated with positive evaluations of the Blackberry (by far the highest probability in that row), negative evaluations of keyboards, are, however, *not* a strong discriminating factor.

For the other entries in the table, we see that criticisms of batteries and the phone network are more associated with negative sentiments towards iPhones.

7.5 DEBATE STANCE CLASSIFICATION

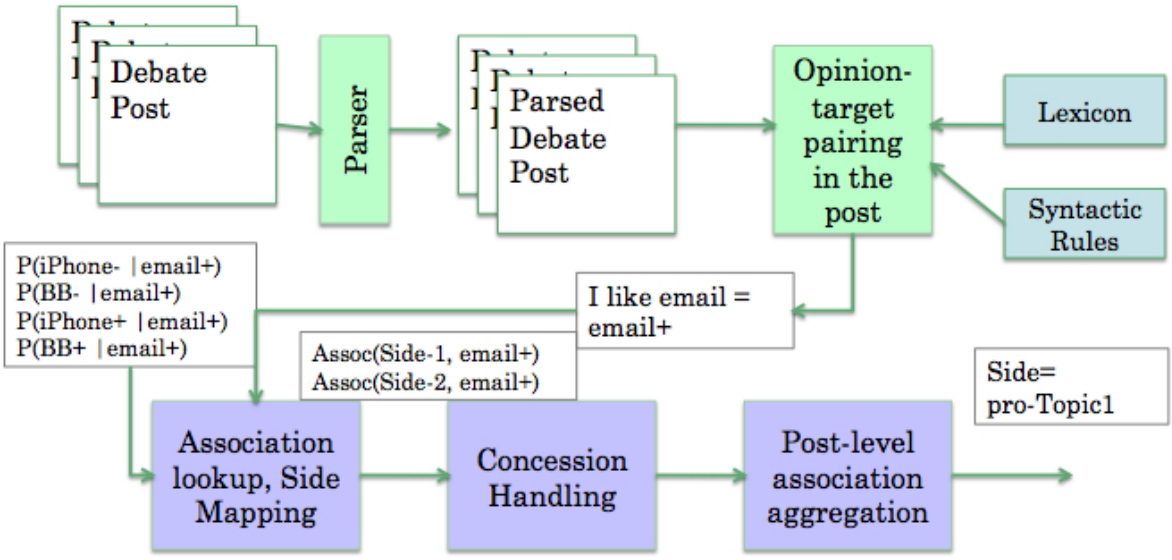

Figure 26: Learning relations from the web

Once we have the probabilities collected from the web (using the procedure from the previous section), we can build our classifier to classify the debate posts. Figure 26 shows the architecture of our system that employs previously learnt associations and classifies the stances of posts.

Here again, we parse the posts and use the process described in Section 7.2 to extract opinion target pairs for each opinion expressed in the post. The green boxes in Figure 26 are the same as that used for web mining. The blue boxes in the figure represent modules used exclusively for stance classification.

We perform side mapping as follows. Let N be the number of instances of opinion target pairs in the post. For each instance I_j ($j = \{1...N\}$), we look up the learned probabilities of Section 7.4 to create two scores, w_j and u_j:

$$w_j = P(topic_1^+|target_i^p) + P(topic_2^-|target_i^p) \tag{7.15}$$

$$u_j = P(topic_1^-|target_i^p) + P(topic_2^+|target_i^p) \tag{7.16}$$

where $target_i^p$ is the opinion-target type of which I_j is an instance.

Score w_j corresponds to $stance_1$ and u_j corresponds to $stance_2$. A point to note is that, if a target word is repeated, and it occurs in different opinion-target instances, it is counted as a separate instance each time — that is, here we account for tokens, not types. Via Equations 7.15 and 7.16, we interpret the observed opinion-target instance I_j in terms of debate sides. We then account for concessionary opinions in the post. The details of this module are explained in Section 7.5.1.

Finally, for post-level aggregation, we formulate the problem of finding the overall side of the post as an Integer Linear Programming (ILP) problem.[4] The side that maximizes the overall side-score for the post, given all the N instances I_j, is chosen by maximizing the objective function

$$\sum_{j=1}^{N}(w_j x_j + u_j y_j) \tag{7.17}$$

[4]Note that even though we use this optimization formulation, a simple aggregation method will work as well.

subject to the following constraints

$$x_j \in \{0, 1\},\ \forall j \tag{7.18}$$

$$y_j \in \{0, 1\},\ \forall j \tag{7.19}$$

$$x_j + y_j = 1,\ \forall j \tag{7.20}$$

$$x_j - x_{j-1} = 0,\ j \in \{2..N\} \tag{7.21}$$

$$y_j - y_{j-1} = 0,\ j \in \{2..N\} \tag{7.22}$$

Equations 7.18 and 7.19 implement binary constraints. Equation 7.20 enforces the constraint that each I_j can belong to exactly one side. Finally, Equations 7.21 and 7.22 ensure that a single side is chosen for the entire post.

7.5.1 Accounting for Concession

As described in Section 7.1.4, debate participants often acknowledge the opinions held by the opposing side. However, they do not endorse these opinions. These opinions are a type of non-reinforcing opinions, in that they do not reinforce the overall stance of the post. In fact, these conceded opinions are indicative of the side the participant really supports.

We recognize concessionary discourse constructs using the Penn Discourse Treebank [Prasad et al., 2007] list of discourse connectives. In particular, we use the list of connectives from the Concession and Contra-expectation category. Examples of connectives in these categories are "while," "nonetheless," "however," and "even if." We use approximations to finding the arguments to the discourse connectives (*ARG1* and *ARG2* in Penn Discourse Treebank terms). If the connective is mid-sentence, the part of the sentence prior to the connective is considered conceded, and the part that follows the connective is considered non-conceded. An example is the second sentence of Example 7.11. If, on the other hand, the connective is sentence-initial, the sentence is split at the first comma that occurs mid sentence. The first part is considered conceded, and the second part is considered non-conceded. The first sentence of Example 7.10 is an instance where this rule applies.

The opinions occurring in the conceded part are interpreted in reverse. That is, the weights corresponding to the sides w_j and u_j are interchanged in equation 7.17. Thus, conceded opinions are effectively made to count towards the opposing side.

7.6 EXPERIMENTS

7.6.1 Data

Our product-based debate data consists of debates downloaded from http://www.convinceme.net. All posts on this site are accompanied by stance annotation (the stances are self-reported by the participants). Figure 27 is a snapshot of the debate website. All posts on the left side support $stance_1$ and all posts on the right side support $stance_2$

We parse the html pages and extract the post content and stance label. We discard posts containing fewer than 5 sentences (this threshold of 5 sentences was arbitrarily chosen). This removes posts that are short replies to previous posts in conversational threads, or posts where participants do not adequately justify their stance.

Our development data consist of three debates, each from a different product domain: iPhone vs. Blackberry (mobile phones domain), xBox vs. Wii (video games domain), Netflix vs. Blockbuster (movie rental domain).

Our test data consists of posts from four debates spanning three domains: Windows vs. Mac (operating systems domain), Firefox vs. Internet Explorer (web browser domain), Firefox vs. Opera (web browser domain), and Sony Ps3 vs. Nintendo Wii (video games domain). This gives a total of 117 posts. We do not need training data as our system is fully unsupervised.

7.6.2 Baseline

For stance classification in dual-topic debates, we cannot have a system based on opinions alone; as mentioned previously, positive and negative opinion expressions are meaningful

Figure 27: Snapshot of the website http://www.convinceme.net

only when we know their targets.

Thus, our baseline, *OpTopic*, also finds opinion target pairs. However, this system considers only explicit mentions of the topic for opinion analysis. For this system, the step of opinion target pairing only finds all $topic_1^+$, $topic_1^-$, $topic_2^+$, $topic_2^-$ instances in the post. For opinion target pairing, it uses the same method as the other systems (the modules shown in green in the Figure 24). However, it does not find target relations or reinforcing associations.

The opinion-topic pairs are counted for each debate stance according to the Equations 7.12 and 7.13. Each post is assigned the side with the higher score.

7.6.3 Results

Performance is measured using the following metrics:

$$Accuracy = \frac{\#Correct}{\#Total\ posts}$$

$$Precision = \frac{\#Correct}{\#guessed}$$

$$Recall = \frac{\#Correct}{\#relevant}$$

and

$$Fmeasure = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

In our task, it is desirable to make a prediction for all the posts; hence $\#relevant = \#Total\ posts$. This results in recall and accuracy being the same. However, all of the systems do not classify a post if the post does not contain the information it needs. Thus, $\#guessed \leq \#Total\ posts$, and precision is not the same as accuracy.

Table 28 reports the performance of four systems on the test data: the OpTopic baseline, the system based on target relations described in Section 7.3 (OpPMI), the system that employs reinforcing relations from the web (Web-asc), and finally the system that employs associations plus handles concessions (Web-asc+Concession).

The results in Table 28 show that OpTopic has low recall. This supports our observation that debate topics are evoked in a variety of ways and only finding opinions towards the main

	OpTopic	OpPMI	Web-asc	Web-asc + Concession
Precision	66.67	56.12	62.83	67.26
Recall	30.77	47.01	60.68	64.96
Fmeasure	42.11	51.16	61.74	66.09

Table 28: Performance of the systems on the test data (117 posts)

topic is not enough. However, this system has a relatively good precision. This is expected, because opinions towards the main topic are very clear in determining the stance.

OpPMI has better recall than OpTopic. Specifically, OpPMI's recall improves over OpTopic by about 17 percentage points, which indicates that it is able to capture the variety of ways in which debate topics are invoked. However, its precision drops by about 10 percentage points. We believe this is due to the addition of noise. Overall, the OpPMI system achieves an improvement of about 9 percentage points in Fmeasure over OpTopic. This result suggests that, target relations are useful improving stance classification, but not all terms that are relevant to a topic are useful for determining the debate side.

Moving on to the system that employs reinforcing associations from the web (Web-asc), we see that it has a pronounced improvement, by about 30 percentage points, in recall over OpTopic. This is accompanied with a drop in precision by about 4 percentage points. This results in an overall improvement in Fmeasure by about 19 percentage points. This indicates that the reinforcing relations we learn via web mining are useful for determining stances.

Finally the system that employs reinforcing associations as well as handles concession (Web-asc + Concession) is the best performer across all metrics. It improves over OpTopic by about 34 percentage points in recall, 0.5 percentage points in precision, and 24 percentage points in Fmeasure. This indicates that our treatment of concessionary opinions is additionally effective.

Comparing the system that learns target relations (OpPMI) to the systems that employ relations between opinion-target pairs (Web-asc and Web-asc+Concession), we see that there

is an improvement in precision by about 6 to 11 percentage points, an improvement in recall by about 13 to 17 percentage points, and an improvement in Fmeasure by 10 to 15 percentage points. This is due to the fact that relations between opinion target pairs are more specific than relations between targets alone.

We performed pairwise comparison of the systems using the McNemar's test. McNemar's test compares how differently the two given systems behave (this is stance classification in our case) for the same data set. We found that OpTopic is significantly different from all systems – the baseline as well as Web-asc and Web-asc+Concession. Web-asc has significantly different classifications from OpTopic and OpPMI. Finally, Web-asc+Concession is significantly different from OpTopic and OpPMI. However, there is no significant difference in the behaviors of the Web-asc and Web-asc+Concession.

Table 29 gives a detailed report of the performance of the four systems for each of the debates in the test data. Note that in three out of four of the debates, the Web-asc+Concession system is able to make a guess for all of the posts (hence, the metrics all have the same values). Also, in three of the four debates, the system using concession handling outperforms the system without it. On average, there is a 5 percentage point improvement in Precision and 5 percentage point improvement in Fmeasure due to the added concession information.

7.7 DISCUSSION

In this section, we qualitatively analyze similarities and differences between the system that employs reinforcing associations (Web-asc) and the system that employs target relations (OpPMI). We then briefly discuss future directions for improvements in our stance recognizers.

	OpTopic	OpPMI	Web-asc	Web-asc+Concession
Firefox Vs Internet explorer (62 posts)				
Prec	67.74	60.0	64.52	66.13
Rec	33.87	53.23	64.52	66.13
F1	45.16	56.41	64.52	66.13
Windows vs. Mac (15 posts)				
Prec	40.0	53.85	66.67	66.67
Rec	13.33	46.67	66.67	66.67
F1	20.0	50.00	66.67	66.67
SonyPs3 vs. Wii (36 posts)				
Prec	80.0	46.15	56.25	68.75
Rec	33.33	33.33	50.0	61.11
F1	47.06	38.71	52.94	64.71
Opera vs. Firefox (4 posts)				
Prec	33.33	100	75.0	100.0
Rec	25.0	50	75.0	100.0
F1	28.57	66.67	75.0	100.0

Table 29: Performance of the systems on the test data

7.7.1 Qualitative Analysis

In order to qualitatively compare the OpPMI system and the Web-asc system, we need to compare Tables 26 and 27 presented in the earlier sections. Table 27 reports the conditional probabilities learned from the web corpus for opinion target pairs used in Web-asc, and Table 26 reports the PMI of the same targets with the debate topics used in OpPMI. First, we observe that the PMI numbers are intuitive, in that all the words, except for "e-mail," show a high PMI relatedness to both topics. All of them are indeed semantically related to the domain.

Next, we see that some conclusions of the OpPMI system are similar to those of the Web-asc system. For example, in Table 26, "Storm" is more closely related to Blackberry than iPhone. In Table 27, positive opinions towards Storm are closely associated with positive opinions towards Blackberry, while negative opinions towards Storm are closely associated with negative opinions towards Blackberry. Thus, both systems have learnt that Storm is *same* as Blackberry, which in this case is a "is-a" relationship.

Similarly, "email" is another case where OpPMI and Web-asc produce the same treatment. First, notice that the PMI measure of "e-mail" suggests that it is not closely related to the debate topics, which again is intuitive, but an undesirable conclusion for the iPhone vs. Blackberry debate. In order to handle such cases OpPMI uses the relative closeness to the two topics; in this case it is closer to Blackberry, and hence a positive opinion towards email denotes a pro-Blackberry stance. The associations learnt by Web-asc also provide the same inference.

However, notice that even though the PMI values for "phone" are intuitive, they may cause errors due to the dual-topic nature of the debates. As iPhone and Blackberry are both phones, the word "phone" does not have a distinguishing power in the iPhone vs. Blackberry debate. That is, $phone^+$ can be equally used to reinforce both $iPhone^+$ and $Blackberry^+$ (as captured by Web-asc, seen in Table 27). Presumably this is because in the discourse, "phone" is used to corefer to a previously mentioned iPhone or Blackberry.

The "network" aspect shows a comparatively greater relatedness to Blackberry than to iPhone. Thus, OpPMI uses it as a proxy for Blackberry. This may be erroneous, how-

ever, because negative opinions towards "network" are more indicative of negative opinions towards iPhones, a fact revealed by Table 27.

In general, the PMI relatedness can only correctly capture the same relation – it learns the degree of closeness of a target to a topic. However, it is not able to distinguish if two terms predominantly co-occur because they are used as alternatives. On the other hand, the Web-asc system can capture alternative relations, even though it does not do this explicitly. For example, given a topic A and a target D, Web-asc learns associations between the following opinion target pairs: A^+–D^+, A^+–D^-, A^-–D^+ and A^-–D^-. Now, suppose if at the end of the web mining phase Web-asc learns that the opinion target pairs A^+ and D^- are closely associated, A^- and D^+ are closely associated (note that "closely associated" implies that there is a reinforcing association), but the opinion target pairs A^+ and D^+ do not have a strong association, nor do the opinion target pairs A^- and D^-. In effect, via these explicit associations, Web-asc has implicitly learnt an alternative relation between A and D; these are associations that would have been observed if A and D were alternatives to each other. Instead of first determining if A and D are alternatives, which would automatically suggest that reinforcing relations between them are represented as either $< +, -, alternative >$(A^+ and D^-) or $< -, +, alternative >$ (A^- and D^+), Web-asc performs inference in the reverse direction.

In fact, by using this approach, Web-asc has more leeway: A^+ and D^- may have a close association, while A^- and D^+ need not. There is no presupposition that both types of reinforcements should exist. Note that this flexibility comes at a cost: establishing the alternative relation between A and D automatically presupposes the two reinforcing associations, while Web-asc has to learn each of them explicitly.

7.7.2 Directions For Future Improvements

7.7.2.1 Removing false lexicon hits We use a comprehensive lexicon that has been vetted in previous research [Wilson et al., 2005b]. However, as shown by [Wiebe and Mihalcea, 2006, Su and Markert, 2008], many subjective words have both objective and subjective senses. Thus, one major source of errors is a false hit of a word in

the lexicon. In order to remedy this, a subjective/objective preprocessing system on lines similar to [Akkaya et al., 2009] can be used.

7.7.2.2 Opinion target pairing The syntactic rule-based opinion target pairing system is a large source of errors in all our systems as well as the baseline. Product review mining work has explored finding opinions with respect to, or in conjunction with, aspects [Hu and Liu, 2004b, Popescu et al., 2005]; however, in our work, we need to find information in the other direction – that is, given the opinion, what is the opinion about.

Improving opinion-target pairing can further improve the performance of all the systems. This can be done by creating more exhaustive rules. The rules could also be evaluated (for precision/recall) against an annotated corpus to identify and remove noisy cases.

7.7.2.3 Pragmatic opinions. Some of our errors are due to the fact that opinions expressed in a post are pragmatic. This becomes a problem especially when a debate post is small, and we have few other lexical clues. The following post is an example:

(7.23) The blackberry is something like $150 and the iPhone is $500. I don't think it's worth it. You could buy a iPod separate and have a boatload of extra money left over.

In this example, the participant mentions the difference in the prices in the first sentence. This sentence implies a negative opinion towards the iPhone ($cost^-$) and a positive opinion towards the Blackberry ($cost^+$). However, recognizing this would require a system to have extensive world knowledge. In the second sentence, the lexicon does hit the word "worth," and, using syntactic rules, we can determine it is negated. However, the opinion target pairing system only tells us that the opinion is tied to "it". A co-reference system would be needed to tie "it" to "iPhone" in the first sentence.

An iterative approach might possibly work for identifying the pragmatic cases. In the initial iterations, the (relatively) clear cases, such as the opinion in the second sentence in the above example, can be resolved. In the later stages, these might be used to infer the pragmatic evaluations.

7.8 RELATED WORK

Other researchers have also mined data to learn associations among products and features. A number of works in product review mining [Hu and Liu, 2004b, Popescu et al., 2005, Kobayashi et al., 2005, Bloom et al., 2007] automatically find features of the reviewed products. However, our approach is novel in that it learns and exploits associations among opinion/polarity, topics, and aspects. In their work on mining opinions in comparative sentences, Ganapathibhotla and Liu [Ganapathibhotla and Liu, 2008] look for user preferences for one product's features over another's. We do not exploit comparative constructs, but rather probabilistic associations. Thus, our approach and theirs are complementary.

Stoyanov and Cardie [Stoyanov and Cardie, 2008b] work on opinion co-reference in the MPQA data; however, we need to identify the specific target.

Several researchers have recognized the important role discourse plays in opinion analysis [Polanyi and Zaenen, 2005, Snyder and Barzilay, 2007, Asher et al., 2008, Kugatsu Sadamitsu and Yamamoto, 2008]. However, previous work did not account for concessions in determining whether an opinion supports one side or the other.

Several researchers have worked on tasks similar to our debate stance classification. Kim and Hovy [Kim and Hovy, 2007] predict the results of an election by analyzing forums discussing the elections. Theirs is a supervised bag-of-words system using unigrams, bigrams, and trigrams as features. In contrast, our approach is unsupervised, and exploits information of discourse-level relations. Bansal et al. [Bansal et al., 2008] and Thomas et al. [Thomas et al., 2006] predict the vote (support/opposition) from congressional floor debates using agreement/disagreement features. We do not model inter-personal exchanges; instead, we model factors that influence stance taking. Lin at al [Lin et al., 2006] identify opposing perspectives. Though apparently related at the task level, perspectives as they define them are not the same as opinions. Their approach does not involve any opinion analysis.

More sophisticated approaches to identifying opinions and recognizing their contextual polarity have been published (e.g., [Wilson et al., 2005a, Ikeda et al., 2008, Kugatsu Sadamitsu and Yamamoto, 2008]). Incorporating these methods into our systems for opinion recognition will improve the results of both, the baseline and our systems.

7.9 SUMMARY

In this work, we perform stance recognition in product debates using opinion analysis. We employ the ideas behind our discourse-level relations to find stances. Under the umbrella of creating a stance recognizer, we create solutions for different subtasks. We explicitly find the targets of opinions, find (a subset of) the *same* target relations, explicitly learn reinforcing relations and handle certain cases of non-reinforcing opinions.

This work takes a leap from the previous chapters in terms of approach, genre and domains. Different from previous chapters, we explicitly find the targets and create opinion-target units. We learn reinforcing relations directly, without depending on the target relations. We create stance classification systems that do not rely on manually annotated data. We use rules for opinion target pairing and concessionary opinions, web-based measures for finding target relations and web mining for learning reinforcing relations. Thus, throughout this work we successfully employ the approach of fully unsupervised systems.

In an attempt to get parser-friendly data and user annotated stances, we ventured into a new genre: online debates. This shift gave us an opportunity to investigate whether our ideas of discourse-level relations manifest in other types of data. The user-reported stances provided a chance for real-word validation of our ideas. By using web data, we were able to test the learning and application of some of our discourse-level relations on different product domains such as phones, browsers, operating systems and video games. The success of our approach in these domains indicates that concepts such as reinforcing and non-reinforcing between opinions, and same and alternative target relations can be learnt and employed in certain other genres and domains.

Our experiments on posts from four product domains reveal that, in general, all of the systems that employ at least some aspects of the discourse-based relations obtain better Fmeasure than baseline. Our system that learns *same* relations between entities (OpPMI) has a drop in precision, but makes up by a large increase in recall, resulting in an overall improvement in Fmeasure. This result supports our hypothesis 7.a that a system that automatically learns and employs elements of target relations can perform better than baseline. Our Web-asc system, that learns reinforcing associations from the web (and in the

process, implicitly learns the *same* and *alternative* relations), gets even more pronounced improvements over the baseline. Finally, the system that learns reinforcing relations plus does concession handling, Web-asc+Concession, has the best precision, recall and Fmeasure, which supports our hypothesis 7.b. This system employs the most of our discourse-based relations. These results in general support our high-level hypothesis that the discourse-level relations introduced in this thesis are useful for determining opinion stances.

As mentioned in Section 7.7.2, the performance of the current systems can be improved by creating a more comprehensive and precise opinion target pairing system, and employing contextual polarity disambiguation.

There are some interesting future directions for this work. This work is on only dual-topic dual-sided debates on named entities. It will be interesting to explore stance classification on multi-topic, multi-sided debates, that is, debates that have more than two sides. We conjecture that in such debates, web mining will be more involved, but essentially an extension of the current approach – as each side may be supported by explicit positive evaluations of the corresponding topic, and also by negative evaluations of either or all of the remaining topics.

8.0 STANCE CLASSIFICATION IN POLITICAL AND IDEOLOGICAL DEBATES

In this chapter, we extend the stance classification work to a more complex type of data: political and ideological[1] debates. Primarily, the complexity arises due to the abstract nature of the topics – the topics are not named entities; rather, they are issues that reveal polarizing ideologies. For example, in the debate "Does God exist", a person may take a *for-existence of God* stance or an *against-existence of God* stance.

We explore how political stances can be recognized using opinion analysis. Building on the observations from the previous chapter, we will use the core idea of opinion-targets as the basic units of information. However, there are a number of challenges in political data that warrant a change in the approach for stance recognition. Preliminary inspection of the development data revealed that the targets of opinions in our debates are usually multiple words representing propositions, situations, etc. that can span over the entire sentence. This makes a word similarity-based approach (for e.g., OpPMI from the previous chapter) for finding target relations unworkable. Hence, we do not find target relations; we only attempt to find relations between opinion target pairs and the debate stances. For instance, an opinion target pair found in the Existence of God debate is $bible^{+}$. We encode this as a feature into a machine learning algorithm which classifies stances (*for-existence of God* or *against-existence of God*)

Ideological stances are based on personal beliefs that can be very specific and variable – a person may choose to believe anything he/she wishes. This makes the idea of mining the web for general associations difficult. Thus, as a first step, we take a supervised approach

[1] In this chapter we work with both political and ideological debates. For brevity, we will use the word "political" or "ideological" interchangeably, to refer to both.

– instead of using web-mining, we use stance-tagged data to learn if and how the opinion-target pairs reinforce a stance. That is, the only associations in this work are between the opinion-targets and the stances. Essentially, the supervised classification approach employs opinion target pairs as features, and learns the associations between these features and the classes (the two stances). For example, the learner may choose to associate $bible^{+}$ with either *for-existence of God* or *against-existence of God*, depending on evidence seen in the training data.

We hypothesize that *by utilizing opinion target pair information, political stances can be recognized better than baseline methods.* Specifically, under this general hypothesis, we have two specific hypotheses:

Hypothesis 8.a A system built using opinion-target information is able to perform better than a baseline method.

Hypothesis 8.b A system built using opinion-target information is able to perform better than a system that uses the current state-of-the-art information.

The availability of many debates for each of our domains makes supervised learning a feasible approach. Political debates are popular on the web. For example, all the following debates "Is there a God", "God does not exist" and "Do you believe in God" serve the same purpose: they elicit the two polarizing stances for the domain of Existence of God. The posts from these debates will convey the same type of information – they will argue for or against the existence of God. We collect posts from such related debates within each domain to give us enough data for our supervised approach. Note that in the product debates, even though we had two debates within the same browser domain (Firefox vs. Opera and Firefox vs. Internet explorer), these debates are unrelated.

Specifically, we have the following specific goals in this chapter:

Exploring more challenging/abstract domains: Our previous work on product debates as well as AMI data dealt with tangible objects such as phones and TV remote controls. Here we explore a more challenging domain – that of political ideology. Our aim is to explore if the basic building blocks such as opinion-targets, that were promising in the product domain, continue to be useful with abstract topics.

Constructing a supervised stance recognizer: As a first step in the exploration of stance classification in more complex domains, we employ a supervised learning framework. Opinion target pairs were the basic units that helped the construction of discourse-level relations in product domains. We encode them as features. Feature analysis of this classifier will provide insights into if and what opinion-targets are useful for reinforcing stances, and will lay the groundwork for exploring more detailed discourse relations in the future.

Creating methods/resources for handling arguing opinions Political debates bring into prominence arguing, a less well explored subjectivity type. Initial data analysis revealed that people employ arguing as much or more than sentiment for debating in these domains. There are no publicly available resources for handling arguing opinions. Thus, we attempt to develop a resource for handling arguing in this work.

The rest of this chapter is organized as follows. We first observe how people express their stances in political debates and discuss the challenges of the domain in Section 8.1. We introduce the data for our experiments in Section 8.2. The construction of the arguing lexicon is explained in Section 8.3, the supervised system is presented in Section 8.4, and the experiments are in Section 8.5. We analyze the different systems in Section 8.6, discuss related work in Section 8.7 and conclude in Section 8.8.

8.1 STANCES IN POLITICAL DEBATES

There are a number of websites that feature political debates. As with our previous data set, the debate posts on these sites have self-reported stance annotation. The political debates discuss controversial issues in domains such as healthcare, gun control, abortion and belief in God. The debate topic is usually a phrase, a proposition or a question. "Universal Healthcare", "All Health Care Should Be Free" and "Should America have Universal Healthcare" are examples of debate topics under the healthcare domain.

For a given domain (e.g. Gay Rights), online political debates present a debate topic, and participants argue for or against the topic. Their stances expressed about the debate topic reflect their stance in the domain. For example, let us consider a specific debate topic

in the domain of Gay Rights, "Should Marriage for Same Sex Couples be Legal?". In this particular debate there are two sides: $side_1$ = Yes and $side_2$ = No. These sides represent two different (polarizing) ideological stances in the domain of gay rights: $stance_1$ = *pro-gay rights* and $stance_2$ = *against-gay rights.* Participants who support the "Yes" side in the "Should Marriage for Same Sex Couples be Legal?" debate belong to the "pro-gay rights" ideology, while the participants who support the "No" side in this debate belong to the "against-gay rights" ideology. Another debate topic in this domain, "Is Homosexuality a Sin?", with $side_1$ = Yes and $side_2$ = No, also reveals the two polarizing ideologies – in this case the participants who support the "Yes" side belong to the against-gay rights stance, while the participants supporting the "No" side belong to the pro-gay rights stance.

To take an example from another domain, Existence of God, the debate titled "Does God really Exist" has two sides: $side_1$ = Yes, belonging to the pro-Existence of God stance, and $side_2$ = No, belonging to the against-Existence of God stance.

Notice that the debate titles in the political debates are not dual-topic, nor are they about named entities. Rather, the titles are propositions or questions about the central topic/issue of the domain (e.g. healthcare, gay rights).

In our new domains, we observe some similarities to debates in the product domains, namely the use of same/alternative aspects, and discourse-level relations. However, these are much more complex and challenging. We also observe that arguing opinions are prominent in these debates. In the following sections, we first illustrate the use of arguing opinions in political debates, and then discuss opinion target pairing and discourse-level opinion relations. Note that even though our observations show that elements of our discourse-level relations manifest in political debates, we do not build systems to capture all of them in this work – this work is rather a first step that employs some of the aspects in a supervised learning framework.

8.1.1 Arguing Opinion

We found that the arguing opinion type is prominent in political debates. In fact, the hot-topic nature of political debates itself presupposes a lot of arguing amongst its participants.

In supporting their side, people argue about what is true (E.g. Existence of God debate) and what should be done (E.g. Healthcare debate). Debate participants express their belief and argue for their side by direct assertions, formulating their ideas as necessary and using emphasis. Examples 8.1, 8.2 and 8.3 illustrate various flavors of arguing opinion expressions in political debates.

(8.1) **Obviously** that hasn't happened, and to be completely objective (as all scientists **should be**) **we must** lean on the side of **greatest evidence** which at the present time is for evolution. [side: against existence of God]

(8.2) So, your argument is essentially that God is a contradiction **precisely because** he cannot contradict himself. This argument is **clearly** logically flawed. [side: for existence of God]

(8.3) I **don't think** Socialism is a dirty word... **but is in fact a necessary** component to a capitalistic society to balance out the **inescapable fact** that capitalism denies the rights of the individuals [side: for healthcare]

In Example 8.1, the participant argues against the existence of God. He uses "**Obviously**" to draw emphasis to his claim. The example also shows the writer arguing for what he believes is right (**should be**) and what is imperative (**we must**). He then uses a superlative construct (**greatest**) to argue for evolution. In Example 8.2, the writer uses the word "**clearly**" to draw attention to his claim that the argument is flawed, and argues by providing reasoning (revealed by the text span **precisely because**). Finally, in the argument for healthcare in Example 8.3, "**don't think**" shows a negative argument that socialism is a dirty word; the **in fact**, **necessary** and **inescapable fact** further bring out the writer's argument against capitalism.

Thus, we need to detect and account for arguing opinions in addition to sentiment opinions in order to capture ideological stances.

8.1.2 Opinion Targets

Targets are vital for determining stances in political debates. Consequently, in this chapter as well, we take an opinion-target approach; that is, aspects of the debate topic are particu-

larized to opinions. However, in political domains, the opinion expressions themselves, and their targets are more complex.

Let us consider the following examples from the healthcare debate. These examples highlight the necessity of the opinion-target approach, and also reveal the complexities involved.

(8.4) Government is a **disease** pretending to be its own cure. [side: against healthcare]

(8.5) ... I **most certainly believe** that there are some **ESSENTIAL, IMPORTANT** things that the government **has or must do** [side: for healthcare]

(8.6) Oh, **the answer is GREEDY** insurance companies that buy your Rep & Senator. [side: for healthcare]

In these debates, we observe that opinions about the debate aspects are important for recognizing the stances. Example 8.4 is a sentence from a post that is against universal healthcare, and it has a negative opinion (sentiment) towards the government (the responsible party for healthcare). On the other hand, posts that support healthcare seem to express positive opinions towards the government (Example 8.5) and negative opinions towards insurance companies (Example 8.6).

Moving on to the complexities, observe that the arguing expressions in the above examples, namely **most certainly believe**, **has or must do** and **the answer is**, are multiple word expressions. This is in contrast to the single word sentiment expressions that were predominant in product debates. Multiple word opinion expressions complicate the opinion target pairing process.

Notice that the targets are more complex too. In the previous chapter, targets were usually single word nouns. This is no longer the case here. In Example 8.5, the entire sentence following the arguing **most certainly believe** is what the participant is arguing about. Similarly, in Example 8.6, the target of the arguing (**the answer is**) is the whole sentence. In general, we found that targets of arguing opinions tend to be clauses or even entire sentences. That is, targets too comprise of *multiple words*.

Target complexity is observed with sentiment expressions too. Often, sentiment opinions seem to affect more than just their immediate targets. In order to see this, let us consider the following examples from the healthcare debate:

(8.7) If there is a right to healthcare, you are **stealing** the provision of that right from someone else.

(8.8) Public education is beset by **exploding** costs, and **deteriorating** quality...

In Example 8.7 we see that the opinion word "steal" indicates that there is a negative opinion towards the "you" (you^-). However, the negative sentiment also applies to the "right to healthcare". In fact the syntactic subject, "you", only partially captures the targets of the negative opinion. Example 8.8 also illustrates this observation. Here, the opinions **exploding** and **deteriorating** are adjectival modifiers of "costs", and "quality" respectively. However, in interpreting the sentence we see that the negativity is in fact directed towards Public education.

These observations indicate that using syntax to tie a single word to the sentiment opinion will not capture all the potential targets. In general, in political debates, we found that opinions often have *multiple targets*.

8.1.3 Target Relations

As with product debates, participants argue for or against a proposition involving the topic via direct reference or by referring to the different aspects associated (or aspects *they* believe are associated) with the topic. Examples 8.4 and 8.5 illustrate this. In Example 8.4, the government is the producer of healthcare, hence the negative opinion towards healthcare is associated with the negative opinion towards the government.

Our political debates have a single main topic. This might suggest that alternative targets will not be as prolific as in the product domain. However, in spite of the apparent single topic nature, the alternative targets do come into play. Example 8.6 is an instance where alternative targets are used. Private insurance companies are generally perceived to be alternatives, or mutually exclusive options to universal healthcare. Hence, in order to argue for universal healthcare, the writer employs his negative evaluation of the insurance companies.

The *same/alternative* relations are prominent even in debates on very abstract topics. For example, in the debate about the existence of God, we observed that science and evidence

are generally perceived to be alternatives to God, while religion and Bible are considered as aspects of God. Examples 8.9, 8.10, 8.11 and 8.12 illustrate this. In Example 8.9, the participant argues for the existence of God directly via positive opinions. The participant in the second example argues for the existence of God too, but in this case it is via negative opinions towards the alternative. In Example 8.11, we see an argument against the existence of God via positive opinions towards science, and in Example 8.12 we see the participant arguing against the existence of God via a direct argument.

(8.9) What **we need to know** is that GOD is **so much much bigger** (beyond our imagination) than what we can think of. [side: for existence of God]

(8.10) Science is all based around theories, but they **may not be true**. [side: for existence of God]

(8.11) Science **doesn't make things up** as it goes along, as religions do. [side: against existence of God]

(8.12) I **am a 100%** atheist because **there is NO** evidence of a God, heaven, hell, devils, angels or a soul that leaves the body.

Table 30 lists some of the observed *same* and *alternative* target relations in the corpus.

Domain	Items in *same* relation	Items in *alternative* relation
Existence of God (main target = God)	Bible, religion, church, creator	science, Darwin, evolution, atheist, genes
Healthcare (main target = healthcare)	socialism, government, Canada, UK	insurance, freedom, liberty, republicans

Table 30: Examples of *same* and *alternative* target relations in our development data

While there is good evidence of the use of the *alternative* and *same* targets, the situation is much more complex and challenging than in product debates. Complexities arise due to the abstract and non-factual nature of ideologies. Specifically, we observe the following:

- *The alternative and same targets are not evoked due to definite, pre-established relations from an ontology.* Rather, these are items/issues *associated* with the topic. For exam-

ple, socialism is not an aspect of healthcare (and was not a relation that the creators of healthcare intended). Rather, it is an issue that is generally associated with it. Additionally, it is not a universal association. It is an association made by a sub-group of people (usually the people who are against healthcare).

- *The associations are valid only in the context of the current debate topic.* For example, within the healthcare debate, positive opinions towards Canada and positive opinions towards healthcare tend to co-occur, and usually negative opinions towards Canada accompany against-healthcare arguments. However, independent of the healthcare debate, it is difficult to find strong associations between Canada and healthcare. This is quite unlike the relations in the product domains, which often hold even outside of the debate. For example, the PMI association (calculated over the entire web) between iPhone and Apple is 0.915, while the association between healthcare and Canada is only 0.51.
- *The relations are due to the personal beliefs and can vary from person to person.* That is, relations between topics and targets (of the *alternative* and *same* categories) are quite variable and vary depending on what people choose to believe. For example, one participant may believe bible and God to be in the *same* relation, another participant need not make this association. Example 8.12 is an interesting instance of this – here the writer groups "God", "heaven", "devil" and "hell" together as items that a theist believes in. He believes this set to be alternative to atheism. However, a theist, or a believer in the supernatural, will consider heaven and hell to be alternatives. Yet more, a whimsical person may choose to make a completely different grouping, and it cannot be considered incorrect, due to the difficulty in establishing what is indeed the ground truth in these cases. Even in relatively more tangible topics such as healthcare, we observe this phenomenon – some people genuinely believe that socialism and universal healthcare are the same while others (equally strongly) believe that they are not. This phenomena is quite unlike the product relations we encountered in the previous chapter where, in most cases, there exists some ground truth regarding the relations (for example, a product and its attributes, or two products made by the same manufacturer).
- *Even within the discourse of the debate, more often than not, items within a set are not related to one another.* For example, we found that in the healthcare debate, people who

are against universal healthcare argue for liberty and are generally sympathetic towards the insurance companies. Consequently, "insurance" and "liberty" belong to the same set – the alternative to universal healthcare, but they are not related to each other directly. This is partly because our debates in this chapter are single-topic and hence elicit "one against all" arguments – people find many (disjoint) alternatives to the central topic. We did not face this complexity in our debates in the previous chapter because they were dual-topic. The alternatives were established by the debate topic itself. Thus, the participants were in some ways subconsciously constrained in the choice of ways to argue for their side.

8.1.4 Relations between opinions

Debate participants express opinions towards various aspects of the debate in order to reinforce their stance. Similar to product debates, due to the goal-oriented nature of the genre, most of the opinions expressed by the participants are meant to reinforce the side they support.

However, we do come across some concessionary opinions. Examples 8.13 and 8.14 illustrate this.

(8.13) That while there are a few **thoughtful, interesting** atheists, the majority are either **rabid** anti-religious **zealots** or people too **ashamed** to take responsibility for their actions in a supernatural sense? [side: for existence of God]

(8.14) Things reproduce to continue their genes, this is a **complicated**, sometimes **counter-intuitive** argument which takes **much explaining** but, once grasped makes **a lot of sense** and **explains a whole mess of problems away** [side: against existence of God]

In Example 8.13, the participant has a positive opinion towards atheists, which he does not fully endorse. The second part of the sentence that reflects his strong negative opinions is the actual prominent opinion that he holds. The writer of the Example 8.14 similarly starts with a negative opinion towards genes (which in this context, is the alternative to God), but then concedes this in favor of the positive opinion that follows.

8.2 DATA FOR POLITICAL DEBATES

Political and ideological debates on hot issues are popular on the web. In this work, we analyze the following domains: Existence of God, Healthcare, Gun Rights, Gay Rights, Abortion and Creationism. Of these, we use the first two for development and the remaining four for experiments and analysis. Each domain is a political/ideological issue and has two polarizing stances: for and against.

Debates within each domain initiate side-taking. The sides eventually map onto the domain-level stances. Table 8.2 lists the domains, examples of debate topics within each domain, the specific sides for each debate topic, and the domain-level stances that correspond to these sides. For example, consider the Existence of God domain in Table 8.2. The two stances in this domain are for-existence of God and against-existence of God. "Do you believe in God", a specific debate topic within this domain, has two sides "Yes!!" and "No!!". The former maps on to for-existence of God stance and the latter maps to against-existence of God stance. The situation is different for the debate "God Does Not Exist". Here, a support for side "against" implies a support for the for-existence of God stance, and a support for side "for" corresponds to a support toward the against-existence of God stance. Similarly, "Should the US adopt a single payer health care program", "Should the US have universal healthcare" and "Is the National Health Care Reform the answer to our medical problems" are three separate debate topics in the healthcare domain. Each of these debates is associated with two stances: "Yes" and "No". In all these cases, the "Yes" sides map onto the for-healthcare stance, and the "No" sides correspond to the against-healthcare stance.

In general we see in Table 8.2 that, the specific debate topic itself may vary, but in each case the two sides for the topic map on to the domain-level stances. We download debates for each domain and manually map the debate-level stance to the stances for the domain. Table 8.2 also reports the number of debates, and the total number of posts for each domain. For instance, we collect 16 different debates in the healthcare domain which gives us a total of 336 posts. All debate posts have self-reported debate-level stance tags. We manually map the debate-level stances to the domain-level (at the level of the main topic) stances. Thus, our corpus consists of posts and their domain-level stance information.

Domain/Topics	$stance_1$	$stance_2$
Healthcare (16 debates, 336 posts)	*for*	*against*
Obama And His Health Care Plan Any Thoughts	for	against
Should the US have universal healthcare	Yes	No
Debate: Public insurance option in US health care	Pro	Con
Healthcare Reform Good Bad or Ugly	I love it!	Bush had it right.
So you still want Nationalized Health Care	Oh, yes please. Thank you.	Hell no!
Existence of God (7 debates, 486 posts)	*for*	*against*
Is there a God	Yes	No
Do you believe in God	Yes!!	No!!
Does God really Exist	Yes	Nope
God Does Not Exist	against	for
God Exists	for	against
Gun Rights (18 debates, 566 posts)	*for*	*against*
People Should Have The Right To Own Guns	for	against
Should Guns Be Illigal	against	for
Does owning a gun make you safer	Yes	No
Would allowing students to carry weapons make college campuses safer	Yes	No
Debate: Right to bear arms in the US	Yes	No
Gay Rights (15 debates, 1186 posts)	*for*	*against*
Are people born gay	Yes	No
Are children with same sex parents at a disadvantage	No	Yes
Is homosexuality a sin	No	Yes
Should marriage for same sex couples be legal	Yes	No
Should california pass prop 8	No	Yes
Abortion (13 debates, 618 posts)	*for*	*against*
Should abortion be legal	Yes	No
Should obama have reversed us abortion policy	Yes	No
Should south dakota pass the abortion ban	No	Yes
Are you Pro Choice	Yes of course	No way
Abortion Should Be Banned	No it should not	Yes it should be
Creationism (15 debates, 729 posts)	*for*	*against*
Evolution Is A False Idea	for	against
Creationism Is Wrong	against	for
Does intelligent design have merit	Yes	No
Has evolution been scientifically proved	It has not	It has
Are religious people who deny evolution stupid	No	Yes
Creationism Or Evolution	Creationism	Evolution

Table 31: Examples of debate topics and their stances

In this work, we use debates from the following websites:[2]

1. http://www.opposingviews.com
2. http://wiki.idebate.org
3. http://www.createdebate.com
4. http://www.forandagainst.com

8.3 ARGUING LEXICON

We observed in Section 8.1.1 that arguing opinions are important in political debates. Arguing is a relatively less explored category in subjectivity. Due to this, there are no available lexicons with arguing terms (clues). Thus, in order to capture instances of arguing we construct an arguing lexicon from an annotated corpus.

The MPQA corpus is annotated for positive and negative arguing expressions. We use this corpus to generate a ngram (up to trigram) arguing lexicon. Specifically, the MPQA corpus (version 2) is annotated with the arguing attitude by Wilson and Wiebe ([Wilson and Wiebe, 2005, Wilson, 2007]). Wilson defines, "Private states in which a person is arguing or expressing a belief about what is true or should be true in his or her view of the world are categorized as Arguing ... The arguing annotations are spans of text expressing the argument or what the argument is". The annotations make the polarity distinction at the category level, that is, there are two arguing categories in the MPQA annotations: *positive arguing* and *negative arguing*.

The examples below illustrate the MPQA arguing annotations (the annotated text spans are shown in bold). Examples 8.15 and 8.18 are examples of positive arguing and Examples 8.16 and 8.17 illustrate negative arguing annotations.

(8.15) Iran **insists its nuclear program is purely for peaceful purposes**.

(8.16) "**It is analogous to the US crackdown on terrorists in Afghanistan**," Ma said

(8.17) Officials in Panama **denied that Mr. Chavez or any of his family members had asked for asylum**.

[2] We do not use data from www.convinceme.net as it was down for the duration of this work.

(8.18) Putin remarked that the events in Chechnia "**could be interpreted only in the context of the struggle against international terrorism**."

Careful inspection of these text spans reveal that the arguing span annotations can be considered to be comprised of two pieces of information. The first piece of information is what we call as the *arguing trigger expression*. The trigger is an indicator that an arguing is taking place, that is, it is the primary component that anchors the arguing annotation. The second component is the expression that reveals more about the argument, and can be considered to be secondary for the purposes of detecting arguing. For instance, in Example 8.15, the annotated span is **insists its nuclear program is purely for peaceful purposes**. Notice here that the word "insists", by itself, conveys enough information to indicate that the speaker is arguing. It is quite likely that a sentence of the form "X insists Y" is going to be an arguing sentence. Thus, in this example, "insists" is the trigger expression, and "its nuclear program is purely for peaceful purposes" is the part revealing more about the argument (the secondary part). Similarly, in Example 8.16 the arguing trigger is the bigram "It is", which conveys an assertion of whatever follows.

On similar lines, in Example 8.17, the negative arguing annotation, **denied that Mr. Chavez or any of his family members had asked for asylum**, can be split into the primary part "denied that" and the secondary part "Mr. Chavez or any of his family members had asked for asylum". In fact, notice that in this particular annotation, there are *two* text spans that can be considered as arguing triggers: the bigram "denied that" and the unigram "denied". Each of these can independently act as arguing triggers (For example in the constructs "X denied that Y" and "X denied Y").

Finally, in Example 8.18 the annotation, **could be interpreted only in the context of the struggle against international terrorism**, has the following independent trigger expressions "could be * only", "could be" and "could". The wild card in the first trigger expression indicates that there could be zero or more words in its place.

Note that MPQA annotations do not provide this primary/secondary distinction. We make this distinction to create general arguing clues such as "insist" and "could". Table 8.3 lists examples of arguing annotations from the MPQA corpus and what we consider as their arguing trigger expressions.

Positive arguing annotations	Trigger Expression
actually reflects israel's determination not to deal with the palestinian issue as one united and indivisible whole	**actually**
am convinced that improving the environment through technological progress can actually enhance our competitiveness and economic growth	**am convinced**
are totally biased to israel	**are totally**
bear witness that mohammed is his messenger	**bear witness**
believe that the polluters are suddenly going to become reasonable	**believe**
can only rise to meet it by making some radical changes	**can only**
has always seen usama bin ladin's hands behind	**has always**
Negative Arguing Annotations	**Trigger Expression**
a gross misstatement of	**a gross**
certainly not a foregone conclusion	**certainly not**
has never been any clearer	**has never**
needs no revision	**needs no**
not too cool for kids	**not too**
rather than issuing a letter of objection to the australian government	**rather than**
there is no explanation for	**there is no**

Table 32: Arguing annotations from the MPQA corpus and their corresponding trigger expressions

Notice that trigger words are generally at the beginning of the annotations. Most of these are unigrams, bigrams or trigrams (though it is possible for these to be longer, as we saw in Example 8.18). Thus, we can create an lexicon of arguing trigger expressions by extracting the starting n-grams from the MPQA annotations.

The process of creating the lexicon is as follows:

1. Generate a *candidateSet* from the annotations in the corpus. Three candidates are generated from the stemmed version of each annotation:
   - The first word
   - The bigram starting at the first word
   - The trigram starting at the first word

   For example, if the annotation is "can only rise to meet it by making some radical changes", the following candidates are extract from it: "can", "can only" and "can only rise".
2. Remove the candidates from the *candidateSet* that are present in the sentiment lexicon (as these are already accounted for in previous research). Specifically, we find that entries with neutral polarity in the sentiment lexicon and many of our candidate unigrams overlap. For example, "actually", which is a trigger word in Table 8.3, is a neutral subjectivity clue in the lexicon.
3. For each candidate in the *candidateSet*, find the likelihood that it is an indicator of positive or negative arguing (that is, it is a reliable arguing trigger) *in the MPQA corpus*. These are likelihoods of the form:

$$P(positive\ arguing|candidate) = \frac{\#candidate\ is\ in\ a\ positive\ arguing\ span}{\#candidate\ is\ in\ the\ corpus}$$

$$P(negative\ arguing|candidate) = \frac{\#candidate\ is\ in\ a\ negative\ arguing\ span}{\#candidate\ is\ in\ the\ corpus}$$

4. Make a lexicon entry for each candidate. This entry contains the stemmed text, and the two probabilities described above.

This process results in an arguing lexicon with 3762 entires, where 3094 entries have $P(positive\ arguing|candidate) > 0$; and 668 entries have $P(negative\ arguing|candidate) > 0$. Table 8.3 lists select interesting expressions from the arguing lexicon. Note that we do not see many unigrams as they are filtered out in the second step of the lexicon creation.

Entries indicative of Positive Arguing ($P(positive\ arguing|candidate) > P(negative\ arguing|candidate)$
be important to, would be better, would need to, be just the, be the true, my opinion, the contrast, show the, prove to be, only if, on the verge, ought to, be most, youve get to, render, manifestation, ironically, once and for, no surprise, overwhelming evidence, its clear, its clear that, it be evident, it be extremely, it be quite, it would therefore
Entries indicative of Negative Arguing ($P(negative\ arguing|candidate) > P(positive\ arguing|candidate)$
be not simply, simply a, but have not, can not imagine, we dont need, we can not do, threat against, ought not, nor will, never again, far from be, would never, not completely, nothing will, inaccurate and, inaccurate and, find no, no time, deny that

Table 33: Examples of positive arguing and negative arguing from the arguing lexicon

8.4 SYSTEM FOR CLASSIFYING POLITICAL STANCES

In this section we explain our methodology for constructing the stance classifier in detail. The complexities discussed in Section 8.1.3 make the idea of web mining for relations difficult. We saw that, even though political debates exhibit the use of *alternative* and *same* target relations, the target relations in ideological debates are variable and very debate-specific. That is, even though a document on the web may contain the words "healthcare" and "republicans" we cannot be guaranteed that they are employed in the *alternative* relation (or even related at all) unless they are used in the same context.

Fortunately for us, political debate websites are prolific, giving us stance-annotated data. Thus, as we have enough tagged data, we can employ supervised learning to employ opinion-target pairs for stance classification directly. That is, we do not attempt to model the associations explicitly; the learner learns the associations between the opinion-targets and the stances directly from the tagged data.

Our focus is thus on feature engineering. The features we use are opinion-target pairs,

which were the units used for finding explicit associations in product debates. We explore two different features sets: features based on arguing opinions and features based on sentiment opinions. These are encoded as binary features into a standard machine learning algorithm.

8.4.1 Arguing-based Features

We create the arguing features primarily from the arguing lexicon constructed in Section 8.3. We construct additional arguing features using modal verbs and syntactic rules. The latter is motivated by the fact that modal verbs such as "must", "should" and "ought" are clear cases of arguing, and are usually involved in simple syntactic patterns with clear targets.

The process for creating features for a post using the arguing lexicon is simple. For each sentence in the post, we first determine if it contains a positive or negative arguing expression by looking for trigram, bigram and unigram matches with the arguing lexicon. If there are multiple arguing expression matches found within a sentence, we determine the most prominent arguing type (whether the sentence is a positive or a negative arguing) by adding up the probabilities of all the positive arguing and the negative arguing expressions in that sentence. Recall that the arguing lexicon provides the positive and negative arguing probabilities for each entry.

Once the prominent arguing type is determined for a sentence, the prefix *ap* (arguing positive) or *an* (arguing negative) is attached to all the content words in that sentence. Thus, in essence, all the content words in the sentence are assumed to be in the target span. Here content words are defined as nouns, verbs, adjectives and adverbs. The arguing feature is denoted as *ap-target* (positive arguing towards *target*) and *an-target* (negative arguing towards *target*). This creates the opinion target pair features for the post. Figure 28 explains the process of constructing the arguing features in detail. Notice here that we prevent the same text span from matching twice. That is, once a trigram match is found, a substring bigram match with the same text span is avoided.

8.4.1.1 Additional Features For Arguing

In addition to arguing features from the lexicon, we construct a small set of additional features that capture clear cases of arguing.

$postArguingFeat \leftarrow \{\ \}$
$positiveArg \leftarrow 0$
$negativeArg \leftarrow 0$
for each stemmed sentence $s_i \in$ post **do**
  $matchedSet \leftarrow \{\ \}$
  for each trigram $t_{ij} \in s_i$ **do**
    if $t_{ij} \in$ arguing lexicon **then**
      $positiveArg \leftarrow positiveArg + P(positive\ arguing|t_{ij})$
      $negativeArg \leftarrow negativeArg + P(negative\ arguing|t_{ij})$
      $matchedSet \leftarrow matchedSet \cup t_{ij}$
    end if
  end for
  for each bigram $b_{ij} \in s_i$ **do**
    $substringMatch \leftarrow false$
    for each $matched \in matchedSet$ **do**
      if b_{ij} is a substring of $matched$ **then**
        $substringMatch \leftarrow true$
      end if
    end for
    if $\neg substringMatch$ & $b_{ij} \in$ arguing lexicon **then**
      $positiveArg \leftarrow positiveArg + P(positive\ arguing|b_{ij})$
      $negativeArg \leftarrow negativeArg + P(negative\ arguing|b_{ij})$
      $matchedSet \leftarrow matchedSet \cup b_{ij}$
    end if
  end for
  for each unigram $u_{ij} \in s_i$ **do**
    $substringMatch \leftarrow false$
    for each $matched \in matchedSet$ **do**
      if u_{ij} is a substring of $matched$ **then**
        $substringMatch \leftarrow true$
      end if
    end for
    if $\neg substringMatch$ & $u_{ij} \in$ arguing lexicon **then**
      $positiveArg \leftarrow positiveArg + P(positive\ arguing|u_{ij})$
      $negativeArg \leftarrow negativeArg + P(negative\ arguing|u_{ij})$
      $matchedSet \leftarrow matchedSet \cup u_{ij}$
    end if
  end for
  if $negativeArg > 0 \mid positiveArg > 0$ **then**
    for each word $w_{ij} \in$ sentence s_i **do**
      if $partOfSpeech\ (\ w_{ij}) \in \{$ Noun, Adj, Adverb, Verb $\}$ **then**
        if $negativeArg \geq positiveArg$ **then**
          $postArguingFeat \leftarrow postArguingFeat \cup concatenate(ap-, w_{ij})$
          else
          $postArguingFeat \leftarrow postArguingFeat \cup concatenate(an-, w_{ij})$
        end if
      end if
    end for
  end if

end for

Figure 28: Creating Arguing features for a post using the Arguing Lexicon

Modals words such as "must", "should" are usually good indicators of arguing. This is a small closed set. Also, the target (what the arguing is about) is syntactically associated with the modal word. Thus, the target of the assertion can be extracted by using a small set of syntactic rules.

We create syntactic rules to capture modal constructs. Figure 29 details the creation of modal features. For every modal detected, three features are created. Note that all the different modals are replaced by "should" while creating features. This helps to create more general features. For example, given a sentence "They must be available to all people", the method creates three features "they should", "should available" and "they should available". These patterns are created independently of the arguing lexicon matches, and added to the feature set for the post.

8.4.2 Sentiment-based Features

Sentiment-based features are created independent of the arguing features. In order to detect sentiment opinions, we use a sentiment lexicon from previous work [Wilson et al., 2005a]. In addition to the positive and negative words, this lexicon also lists sentiment words that are themselves neutral with respect to polarity. Examples of neutral entries are "absolutely", "amplify", "believe", "surprise" and "think". While these were not very significant for product domains, they seem important in political domains, as the neutral subjective words draw attention or emphasis to their targets. Thus, in this work, we have three polarities of sentiments: positive ($^{+}$), negative ($^{-}$) and neutral ($^{=}$).

We observed in Section 8.1.2 that the sentiment opinions affect words throughout the sentence. To account for this, we find the sentiment polarity of the entire sentence and assign this polarity to each content word in the sentence. In order to detect the sentence polarity using the sentiment lexicon, we use the Vote and Flip algorithm (Figure 8.4.2) from Choi and Cardie [Choi and Cardie, 2009]. The algorithm essentially counts the number of positive, negative and neutral lexicon hits in a given expression and accounts for negator words. The algorithm is used as is, except for the default polarity assignment (as we do no know the most prominent polarity in the corpus – this may vary according to the debate

$postModalFeat \leftarrow \{\ \}$

$MODALS \leftarrow \{$ should, would, must, could, will, ought, might, can, shall $\}$

for each stemmed sentence $s_i \in$ post **do**

  for each word $w_{ij} \in$ sentence s_i **do**

    if $w_{ij} \in MODALS$ **then**

      $modal \leftarrow$ "$should$"

      if w_{ij} is modified by negation **then**

        $modal \leftarrow$ "$shouldnot$"

      end if

      $par \leftarrow$ parent of w_{ij} in the dependency tree

      $subj \leftarrow$ word functioning as the nominal subject of par in the dependency tree

      $obj \leftarrow$ word functioning as the direct object of par in the dependency tree

      $postModalFeat \leftarrow postModalFeat \cup concatenate(subj, modal)$

      $postModalFeat \leftarrow postModalFeat \cup concatenate(modal, obj)$

      $postModalFeat \leftarrow postModalFeat \cup concatenate(subj, modal, obj)$

    end if

  end for

end for

Figure 29: Creating Arguing features using Modals and syntactic rules

topic). Note that the Vote and Flip algorithm has been developed for expressions but we employ it for the whole sentence.

Once the polarity of a sentence is determined, we create sentiment features for the sentence. This is done for all sentences in the post. The process for creating sentiment features for debate posts is described in Figure 31.

8.5 EXPERIMENTS

Our experiments are carried out on debate posts from the following four domains: Gun Rights, Gay Rights, Abortion, Creationism. We create a corpus with equal class distribution for each domain. This is done as follows: we merge all debates from a domain and sample instances (posts) from the majority class to obtain equal numbers of instances for each stance. This gives us a total of 306 posts for the Gun Rights domain, 846 posts for the Gay Rights domain, 550 posts for the Abortion domain and 530 posts for the Creationism domain.

Our first baseline is a distribution-based baseline which has an accuracy of 50%. We also construct *Unigram*, a system based on unigram content information, that is, there is no explicit treatment for opinion polarity identification. Unigrams are very reliable for stance classification in the political domain (and have been effectively employed in previous work, for e.g., [Lin et al., 2006, Kim and Hovy, 2007, Thomas et al., 2006]). Preliminary inspection of our data also gave similar insights. Evoking a particular topic can be very indicative of a stance. For example, a participant who uses the words "child" and "life" in an abortion debate is more likely from an against-abortion side; while the choice of words such as "woman", "rape" and "choice" indicate a for-abortion stance. Hence, in political debates, the issues that participants chooses to address are in themselves very indicative of the stances.

We construct three systems that use opinion information: The *Sentiment* system that uses only the sentiment features described in Section 8.4.2, the *Arguing* system that uses only arguing features constructed in Section 8.4.1, and the *Arg+Sent* system that uses both sentiment and arguing features for each post.

for each expression e_i **do**

nPositive $\leftarrow$ # of positive words in e_i

nNeutral $\leftarrow$ # of neutral words in e_i

nNegative = # of negative words in e_i

nNegator = # of negating words in e_i

if (nNegator % 2 = 0) **then**

fFlipPolarity $\leftarrow$ false

else

fFlipPolarity $\leftarrow$ true

end if

end for

if (nPositive > nNegative) & $\neg$ fFlipPolarity **then**

Polarity(e_i) $\leftarrow$ positive

else if (nPositive > nNegative) & fFlipPolarity **then**

Polarity(e_i) $\leftarrow$ negative

else if (nPositive < nNegative) & $\neg$ fFlipPolarity **then**

Polarity(e_i) $\leftarrow$ negative

else if (nPositive < nNegative) & fFlipPolarity **then**

Polarity(e_i) $\leftarrow$ neutral

else if nNeutral > 0 **then**

Polarity(e_i) $\leftarrow$ neutral

else

Polarity(e_i) $\leftarrow$ default polarity (the most prominent polarity in the corpus)

end if

Figure 30: Choi and Cardie's Vote and Flip algorithm

postSentimentFeat ← { }
for each sentence $s_i \in$ post **do**
  pol ← $Vote - flip$(sentence)
  if pol ≠ null **then**
    for each word $w_{ij} \in$ sentence s_i **do**
      if $partOfSpeech$ (w_{ij}) ∈ { Noun, Adj, Adverb, Verb } **then**
        postSentimentFeat $\bigcup$ w_{ij}^{pol})
      end if
    end for
  end if
end for

Figure 31: Creating sentiment features for a post

All systems are implemented using the standard implementation of the SVM classifier in the Weka toolkit [Hall et al., 2009] (weka.classifiers.functions.SMO). We measure the performance using the accuracy metric. In the supervised setting, all instances are classified. Thus, the total number of instances (the denominator in accuracy calculation) is the same as the number of instances to be classified (the number of relevant instances; the denominator for recall calculation), which is also the same as the number of instances each classifier classifies (the number of retrieved instances; the denominator for precision calculation). Hence, the accuracy is the same as the overall precision and recall metrics (as precision and recall are the same, the Fmeasure will also have the same numeric value). Thus, we report only the classification accuracy.

Table 34 shows the accuracy averaged over 10 fold cross-validation experiments for each domain. The first row (Overall) reports the accuracy over all the 2232 posts in the data.

Overall, we notice that all the supervised systems perform better than the distribution-based baseline. Observe that Unigram has a better overall performance than Sentiment, while

Domain (#posts)	Distribution	Unigram	Sentiment	Arguing	Arg+Sent
Overall (2232)	50	62.50	55.02	62.59	63.93
Guns Rights (306)	50	66.67	58.82	69.28	70.59
Gay Rights (846)	50	61.70	52.84	62.05	63.71
Abortion (550)	50	59.1	54.73	59.46	60.55
Creationism (530)	50	64.91	56.60	62.83	63.96

Table 34: Accuracy of the different systems

its performance is almost the same as Arguing. The good performance of the Unigram system indicates that the choice of words, what participants choose to speak about, is a good indicator of ideological stance taking. This result confirms previous researchers' intuition that, in general, political orientation is a function of "authors' attitudes over multiple issues rather than positive or negative sentiment with respect to a single issue" [Pang and Lee, 2008]. Nevertheless, the Arg+Sent system that uses both arguing and sentiment features outperforms Unigram.

We performed McNemar's test to measure the difference in system behaviors. The test was performed on all pairs of the supervised systems using all 2232 posts. The results show that there is a significant difference between the classification behavior of Unigram and Arg+Sent systems ($p < 0.05$). The difference between classifications of Unigram and Arguing approaches significance ($p < 0.1$). There is no significant difference in the behaviors of all other system pairs.

Moving on to the detailed performance in each domain, we see that Unigram outperforms Sentiment for all domains. Arguing and Arg+Sent outperform Unigram for three domains (Guns, Gay Rights and Abortion), while the situation is reversed for one domain (Creationism). We carried out separate t-tests for each domain, using the results from each test fold as a data point. Our results indicate that the performance of Sentiment is significantly different from all other systems for all domains. However there is no significant difference between

the performance of the remaining systems.

8.6 DISCUSSION

Unigram System
Attributes corresponding to for-gay rights
jake(0.6061) sleep(0.5327) understand(0.489) single(0.4449) equal(0.4322) its(0.4259) sapiens(0.4211) sinner(0.3899) prior(0.3883) day(0.3875) be(0.3825) basically(0.3579) joe(0.3534) interracial(0.3518) country(0.3477) progress(0.3449) ever(0.3441) fight(0.3439) hahahaha(0.3352) second(0.3333)
Attributes corresponding to against-gay rights
nooooooooooooooo(-0.4896) spend(-0.4133) agency(-0.3849) curious(-0.3597) vote(-0.3458) hand(-0.3451) protect(-0.3352) place(-0.3229) -rrb-(-0.3193) alot(-0.3071) disgusting(-0.3032) lot(-0.3009) arrogant(-0.3003) open(-0.2986) definition(-0.2972) circumstance(-0.2929) sex(-0.292) blame(-0.2916) purpose(-0.2892) shun(-0.2851)

Table 35: Top attributes in Unigram

The performance measured over the four political domains indicate that, overall, Unigram is better than the sentiment-based system, and Arg+Sent is better than the unigram-based system. Unigram essentially captures what the participants are writing about, while opinion-based systems capture what the opinions are about. Due to the hot-topic nature of the debates, most of the sentences express arguing opinions. Thus, there is a large overlap between the top features selected by the classifier in the Unigram, Arg+Sent and Arguing systems.

Inspection of the top features used for discriminating the classes seems to corroborate this. Tables 35, 36, 37 and 38, respectively, present the top features used by Unigram, Arg+Sent, Arguing and Sentiment systems for the Gay Rights domain. The first row of

Arg+Sent System
Attributes corresponding to for-gay rights
ap-understand(0.3043) ap-single(0.2998) ap-gay(0.289) ap-be(0.287) do^{-}(0.2866) ap-heterosexual(0.2746) ap-jake(0.271) ap-equal(0.2666) ap-u(0.2506) ap-straight(0.2399) ap-do(0.2239) ap-here(0.2176) ap-religious(0.216) be$^{=}$(0.2143) ap-country(0.2066) an-jake(0.1984) same$^{=}$(0.1982) think$^{=}$(0.1964) ap-need(0.1929) ap-family(0.1879)
Attributes corresponding to against-gay rights
an-nooooooooooooo(-0.4126) ap-god(-0.3029) ap-i(-0.2615) should-get(-0.2561) ap-sin(-0.2371) an–rrb-(-0.2292) ap-now(-0.2147) ap-sex(-0.2097) ap-hand(-0.2) ap-opposite(-0.1921) ap-spend(-0.1889) ap-im(-0.1863) ap-mind(-0.1777) people-should(-0.168) ap-look(-0.1669) ap-christ(-0.1661) person$^{=}$(-0.1655) ap-woman(-0.1653) ap-move(-0.1644) ap-someone(-0.1636)

Table 36: Top attributes in Arg+Sent

Arguing System
Attributes corresponding to for-gay rights
ap-equal(0.373) ap-understand(0.3725) ap-single(0.338) ap-jake(0.3273) ap-be(0.2972) ap-religious(0.2968) ap-gay(0.292) ap-u(0.2848) ap-sleep(0.2819) an-jake(0.2701) ap-here(0.2687) ap-country(0.2616) ap-above(0.2616) ap-straight(0.2591) ap-do(0.257) ap-progress(0.2472) ap-wait(0.2423) ap-hetero(0.2405) ap-heterosexual(0.2399) ap-group(0.2395)
Attributes corresponding to against-gay rights
an-noooooo(-0.501) ap-sin(-0.3586) ap-i(-0.3205) ap-god(-0.3066) should-get(-0.304) an-?(-0.298) an–rrb-(-0.2942) ap-sex(-0.2846) ap-now(-0.2436) ap-christ(-0.2396) ap-alot(-0.2352) ap-spend(-0.2271) ap-disgusting(-0.2156) ap-vote(-0.2142) ap-hand(-0.2111) ap-im(-0.2075) ap-calling(-0.2058) ap-purpose(-0.2029) ap-mind(-0.2024) couple-should-say(-0.2023)

Table 37: Top attributes in Arguing

Sentiment System
Attributes corresponding to for-gay rights
congratulation^{+}(0.9856) anyway$^{=}$(0.9853) video^{+}(0.9642) meaning^{+}(0.8979) clear^{+}(0.875) table$^{=}$(0.8623) change^{+}(0.8567) trivial^{-}(0.8496) harm$^{=}$(0.8476) session^{+}(0.8394) hurt$^{=}$(0.8116) forget^{-}(0.809) way^{+}(0.8063) comprehend^{+}(0.806) screw^{-}(0.7916) anything^{-}(0.791) million$^{=}$(0.7906) wonder$^{=}$(0.782) hurt^{-}(0.7816) something^{-}(0.7724) ok$^{=}$(0.7668) discrimination^{-}(0.7643) assessment^{+}(0.7637) incorrect$^{=}$(0.7582) pointless$^{=}$(0.7574) different$^{=}$(0.7571) class^{-}(0.7536) debate^{+}(0.7451) nothing^{+}(0.7436) child-abuse^{-}(0.7417) error^{-}(0.7382) try^{-}(0.7374) stigma$^{=}$(0.733) post^{+}(0.7237) word^{-}(0.7208) lesbian^{+}(0.7201)
Attributes corresponding to against-gay rights
person$^{=}$(-0.9931) conviction^{+}(-0.945) lie^{-}(-0.9334) ponder$^{=}$(-0.9093) heathen^{-}(-0.899) rebuttal^{+}(-0.8971) grateful^{+}(-0.8954) good^{+}(-0.8754) affect^{+}(-0.8677) scare^{-}(-0.8608) sin^{-}(-0.8599) life^{+}(-0.8503) man^{+}(-0.8186) christ$^{=}$(-0.8012) oppose^{-}(-0.7959) logic^{+}(-0.7869) first-hand$^{=}$(-0.7804) thank^{+}(-0.7639) funny^{+}(-0.7511) rights^{-}(-0.7463) loose$^{=}$(-0.7263) beautiful^{+}(-0.7183) curious^{+}(-0.7181) hope^{+}(-0.7179) well$^{=}$(-0.7075) purpose$^{=}$(-0.6931) animal^{+}(-0.6882) minded$^{=}$(-0.687) deny$^{=}$(-0.6844) only$^{=}$(-0.6676) minority$^{=}$(-0.662) arrogant$^{=}$(-0.6591) opinion^{+}(-0.647) question$^{=}$(-0.6289) truth^{+}(-0.6284) kids$^{=}$(-0.6212) picture$^{=}$(-0.6212) say^{-}(-0.6145) family$^{=}$(-0.6129) scripture$^{=}$(-0.6082) hopfully^{+}(-0.6048) survival^{+}(-0.6026)

Table 38: Top attributes in Sentiment

each table lists the features useful for the for-gay rights stance and the second row lists the features useful for the against-gay-rights stance. The attribute weights assigned by the SVM classifier is shown in braces.

In Table 35 words such as "understand", "sinner", "country", "progress", "religion" appear important for arguing for gay rights, while words such as "disgusting", "arrogant", "definition", "purpose", "shun" and "god" are important indicators of against-gay rights stance. Presumably, the pro-gay rights posts highlight the need for progress and understanding, while the against gay rights posts highlight the definition and purpose of marriage. These features are somewhat intuitive: the definition of marriage and the arguments involving God are important for people against gay marriage while progress and understanding is valued more by people arguing for it.

Observe there is an overlap between the words in Table 35 and Table 36. We see in Table 36, that pro-gay rights people argue for "understand", "equal", "gay", while against-gay rights people argue for "god", "christ", "opposite" and "disgusting". Notice that almost all features for both sides are with positive polarities: either positive arguing or positive sentiments. This indicates that people choose to support their stances by talking about things that matter to them the most; they highlight what they value. The main reason why we do not see negative arguing in the table is that there are fewer negative arguing entries in the lexicon as compared to positive arguing. The sentiment lexicon does have a large number of negative polarity entries, but the sentiment features in general are not as useful as the arguing features, as is evident from the table. Also, we have observed in our development data that participants have a preference to emphasize what they consider important justification for their stance. For example, in arguing against the existence of God, participants are less likely to explicitly say bad things about God. Rather, they say good things about science and evidence.

Table 36 also provides an intuition into the relatively lower performance of Sentiment as compared to Arguing and Arg+Sent – even when both types of features are used, the features considered most useful by the classifier are of the arguing type, which indicate that arguing opinions are more important for stance classification in political debates.

Arguing captures similar arguing information as Arg+Sent and consequently there is

a large overlap between the entries in Table 36 and 37. The classifier that relies only on sentiment opinions (Sentiment) shows a different set of features that are considered useful. Only some of the top features listed in Table 38 are intuitive (note that the neutral opinions are shown as $^{=}$). Pro-gay rights posts express positive sentiments towards "lesbian" and negative sentiment towards "discrimination". The against-gay rights posts speak positively about "survival" and negatively about "rights".

Unigram Features	
For Gay Rights	Against Gay Rights
constitution, fundamental, rights, suffrage, pursuit, discrimination, government, happiness, shame, wed, gay, heterosexuality, chromosome, evolution, genetic, christianity, mormonism, corinthians, procreate, adopt	pervert, hormone, liberty, fidelity, naval, retarded, orientation, private, partner, kingdom, bible, sin, bigot

Table 39: Examples of unigram features associated with the stances in Gay Rights domain

The overlap between the top features in Unigram and Arg+Sent explains why there is no pronounced difference between the overall performance of the two systems in political debates. However, we believe that Arg+Sent makes a finer distinction based on the polarity of the opinions towards the same set of words. Tables 39 and 40 list some interesting features belonging to the two ideological stances in the Gay Rights domain for the two classifiers. Table 39 shows the words corresponding to the stances in the Unigram classifier. These are indeed intuitive – "fundamental", " rights", "discrimination" "evolution" and "christianity" are important terms for for-gay rights, while "pervert", "orientation", "sin" and "bible" are useful for arguing against gay rights. However, it is not evident if a word is evoked as, for example, a pitch, concern, or denial. Also, notice that related terms such as "christianity" and "bible" belong to opposite sides – "bible" is an indicator for against-gay rights and "christianity" is an indicator for for-gay rights, which is confusing.

The arguing features shown in Table 40 seem to partially resolve these ambiguities. Here

the separation between features is even more informative and intuitive – positive arguing about "christianity", "corinthians", "mormonism" and "bible" are all indicative of against-gay rights stance. These are indeed the beliefs and concerns that shape the against-gay rights stance. Notice that there is negative arguing associated with each of the above mentioned words and these are features indicating a for-gay rights stance. Presumably, these occur in refutations of the concerns influencing the opposite side. Similarly, positive arguing about terms such as "genetics", "orientation" and "chromosome" represent the arguments from the pro-gay community about homosexuality not being a life-style choice. Likewise, the appeal for equal rights for gays is captured by the features that show positive arguing about "liberty", "independence", "pursuit" and "suffrage". On the other hand, a negative arguing associated with "pervert" reveals the protest by the pro-gay community about considering gays abnormal.

In general, unigram features associate the choice of topics (and words) with the stances, while the arguing features can capture the concerns, defenses, appeals, refutations or denials that signify each side. In the current work, we do not explicitly make this fine-grained distinctions. Information regarding these categories can be added in the lexicon in the future, if such distinctions are required.

Interestingly, the opinion target pairs in Table 40 reveal an implicit demarcation of *same* and *alternative* targets. Both positive arguing for "gay" and negative arguing involving "heterosexuality" support a for-gay rights stance – these are mutually exclusive options in the context of the discourse. Similarly, each stance is associated with opinions of like polarity towards items that are essentially the same. For instance the for-gay rights stance is associated with positive arguing about "rights", "liberty", "suffrage", "independence" and negative arguing about "christianity", "mormonism" and "corinthians".

Table 41 lists a few examples of sentiment-based features from the Arg+Sent classifier. Some of these features are intuitive – the features indicate that the pro-gay rights arguments have a positive sentiment regarding "rights", "freedom" and "gay" and negative sentiments regarding "ban" and "faith". However, we found that many of the sentiment features for the classifier are not as intuitive and informative as the arguing features discussed above.

Arguing Features			
For Gay Rights	Against Gay Rights	For Gay Rights	Against Gay Rights
ap-constitution	an-constitution	ap-fundamental	an-fundamental
ap-rights	an-rights	ap-hormone	an-hormone
ap-liberty	an-liberty	ap-independence	an-independence
ap-suffrage	an-suffrage	ap-pursuit	an-pursuit
ap-discrimination	an-discrimination	an-government	ap-government
ap-fidelity	an-fidelity	ap-happiness	an-happiness
an-pervert	ap-pervert	an-naval	ap-naval
an-retarded	ap-retarded	an-orientation	ap-orientation
an-shame	ap-shame	ap-private	an-private
ap-wed	an-wed	ap-gay	an-gay
an-heterosexuality	ap-heterosexuality	ap-partner	an-partner
ap-chromosome	an-chromosome	ap-evolution	an-evolution
ap-genetic	an-genetic	an-kingdom	ap-kingdom
an-christianity	ap-christianity	an-mormonism	ap-mormonism
an-corinthians	ap-corinthians	an-bible	ap-bible
an-sin	ap-sin	an-bigot	ap-bigot
an-procreate	ap-procreate	ap-adopt	an-adopt

Table 40: Examples of arguing features from Arg+Sent that associated with the stances in the Gay Rights domain

Sentiment Features			
For Gay Rights	Against Gay Rights	For Gay Rights	Against Gay Rights
constitution$^+$	constitution$^-$	rights$^+$	rights$^-$
gay$^+$	gay$^-$	adopt$^+$	adopt$^-$
hormone$^-$	hormone$^+$	orientation$^-$	orientation$^+$
vote$^-$	vote$^+$	assumption$^-$	assumption$^+$
choice$^+$	choice$^-$	faith$^-$	faith$^+$
evidence$^-$	evidence$^+$	ban$^-$	ban$^+$
marry$^+$	marry$^-$	freedom$^+$	freedom$^-$

Table 41: Examples of sentiment features from Arg+Sent that associated with the stances in the Gay Rights domain

8.7 RELATED WORK

Generally, identifying political viewpoints is considered different from sentiment analysis, and has employed information from words in the document, instead of using opinion information. Unigram information has been employed for classifying political leanings [Malouf and Mullen, 2008, Mullen and Malouf, 2006, Grefenstette et al., 2004, Laver et al., 2003, Martin and Vanberg, 2008], and document perspectives (whether a given document is written from an Israeli or Palestinian perspective) [Lin et al., 2006, Lin and Hauptmann, 2006, Lin, 2006]. Specifically, Lin et al. observe that people from opposing perspectives seem to use words in differing frequencies. Additionally, researchers have also used words in the document to classify support/opposition in congressional floor debates [Thomas et al., 2006, Bansal et al., 2008]. On similar lines, Kim and Hovy [Kim and Hovy, 2007] use unigrams, bigrams and trigrams for election prediction from forum posts.

In contrast, our work specifically employs sentiment-based and arguing-based features

to perform stance classification in political debates. Our experiments are focused on determining *how the different opinion expressions reinforce an overall political stance*. Our Unigram system uses the same type of information as the works listed above. Results from our work indicate that while the unigram information is reliable, further improvements can be achieved in certain domains using our opinion-based approach.

Another difference from previous work on political/ideological classification is that the topics we explore are different.

Discourse-level participant relation, that is, whether participants agree/disagree has been found useful for determining political side-taking. For example, researchers have observed that responses to forum posts are predominantly due to disagreements [Agrawal et al., 2003]. Thus, two adjacent posts in a thread are likely to belong to opposite political ideologies. Similarly researchers have found that quotations are useful indicators of ideological disagreements [Malouf and Mullen, 2008]. Consequently, previous research [Thomas et al., 2006, Bansal et al., 2008] has focussed on explicitly classifying agreement/disagreement between participants in congressional floor debates and utilizing this in a min-cut framework to improve stance classification. Agreement/Disagreement relations are not the main focus of our work. Our work is concerned with opinion analysis, and specifically on how opinions reinforce a particular stance. We believe our work is complementary to the discourse-level agreement/disagreement approach, and incorporating both can further improve performance.

Other work in the area of polarizing political discourse analyze co-citations [Efron, 2004] and linking patterns [Adamic and Glance, 2005]. In contrast, our focus is on document content and opinion expressions.

In this work, we found that the arguing category is useful for stance classification in political and ideological debates. This is similar to the conclusion in our previous work [Somasundaran et al., 2007b], where we found that making a fine-grained distinction between arguing and sentiment categories is useful for an opinion QA task. When arguing type questions (for e.g. "Should Iran be referred to the security council?") are matched with arguing type answers, performance of the QA system is improved.

Wilson [Wilson, 2007] constructs an automatic system for recognizing arguing expressions in the MPQA corpus. However, she uses the same set of words as the sentiment lexicon from

[Wilson et al., 2005a], with manually added information of whether it indicates positive or negative arguing. Only 2.6% of entries from the sentiment lexicon were marked as having a positive arguing polarity, and even fewer, 1.8%, were marked as having a negative arguing polarity. In contrast, we create an arguing lexicon automatically from an annotated corpus; we remove overlaps between our arguing lexicon and the sentiment lexicon. Thus, the clues used by our Arguing system do not have any overlap with the clues used in the Sentiment system.

Arguing and categories similar to arguing have been studied in linguistics: Biber [Biber, 1988] in work on textual variation identifies a dimension of "Overt persuasion" whose categories (e.g. modal verbs and conditionals) are similar to the arguing expressions found in our corpus. Ducrot [Ducrot, 1973] studies arguing related items, but his work is on French and is not corpus-based. A vast body of work exists within linguistics, rhetoric and philosophy that is relevant to arguing (e.g., [Dancygier, 2006, van Eeemeren and Grootendorst, 2004]).

8.8 SUMMARY AND FUTURE WORK

In this chapter we explored how opinion target pairs reinforce an overall stance.

We found that debates in political domains are very different from debates in product domains. Unlike product debates, political debate titles are abstract propositions and questions. Also, ideological arguments are complex, quite variable, and dependent on the debate context. Thus, as an exploratory step, we formulated the stance recognition as a supervised classification problem.

We used opinion target pairs as features and the system learns which features are indicative of the stances, based on the evidence in the data. Our supervised approach is facilitated by the copious availability of political discussions on the web.

Our initial observations revealed that, what a debate participant choses to speak about is a strong indicator of his/her stance. We also noticed that arguing, a less well explored category of subjectivity, is prominent in political argumentation. We created an arguing

lexicon from the MPQA annotations in order to capture these cases.

We also observed that, unlike in product debates, targets in political debates are more distributed throughout the sentence. Also, targets are often composed of multiple words. In order to account for this, we modified the opinion target pairing strategy.

We performed supervised learning experiments on four different domains. Our results show that both unigram-based and opinion-based systems are useful for improving over baseline methods. This proves our Hypothesis 8.a that the opinion-based systems perform better than a distribution based baseline.

However, our results do not clearly support the second hypothesis (Hypothesis 8.b). Specifically, even though the sentiment-based system is able to perform better than the distribution based baseline, it does not perform at par with the unigram system. Overall, our arguing-based system does as well as the unigram-based system and our system that uses both arguing and sentiment features obtains a marginal improvement over this. However, as these improvements do not show statistical significance, the hypothesis lacks support.

Our future work will focus on improving the performance of our systems. Our policy of tying the prominent sentence-level sentiment and arguing to all the content words was necessary to increase the recall, but this also adds noise, especially when there are multiple opinions of different polarities in a sentence. In future, we plan to explore sub-sentential spans (for example, clauses) in order to make the opinion-target pairing more precise.

Our feature analysis indicates that there is an overlap between the top unigram and arguing features. However, in general, we observed that the arguing features are more informative, as they make finer distinctions that reveal the underlying ideologies. For example, we saw that arguing features in the Gay Rights domain reveal beliefs, accusations, peeves, denials and remonstrations, while the unigram features reveal only topics of importance. In the future, we will work on identifying these specific subtypes of arguing. While the fine-grained distinction in our Arg+Sent system does not produce a significant impact on post-level stance classification, this distinction would be useful for say, a question-answering system that has to specifically find the details associated with stance taking. Our future work will explore ways in which the information revealed via our opinions can be better employed for applications.

There are a number of other avenues to pursue for future work. Concessionary constructs that were easily handled with rules in product debates were not addressed in this work. This was in part due to the supervised formulation, and the fact that the sentences are more complex and the targets are more distributed throughout the sentence. We believe that handling concessions and reducing noise due to these cases will further improve the performance of our Arg+Sent system. Thus, in the future we would like to explore ways to harness concessionary constructs in political debates.

Another direction for future exploration is to improve the precision and recall of the arguing lexicon. The lexicon used in this chapter is created from the annotations in the MPQA corpus. We do not perform fine tuning of the entries – any entry that has a probability score of greater than zero is considered as an arguing clue. In the future, we would like to add more information to the lexicon and explore probability thresholds for improving precision. In order to improve recall, we will explore regular expressions in addition to the current n-gram entries.

9.0 CONCLUSIONS AND FUTURE DIRECTIONS

Our work endeavors to push opinion analysis into new challenging directions in discourse-level analysis. This dissertation follows a research cycle exploring the idea of discourse-level relations for opinion analysis. In order to study the discourse-level relations we design a linguistic scheme to capture these relations. Manual annotation studies are carried out to test human reliability for recognizing these relations and to create annotated data for supervised learning experiments. We utilize these annotations to improve fine-grained polarity recognition. By employing linguistically motivated features and global classification paradigms, we attempt to learn the discourse-level relations. We also employ the ideas from our linguistic scheme to real world data. Via web mining and rules, discourse-level relations are learnt and employed for improving stance recognition in product debates.

9.1 SUMMARY OF CONTRIBUTION AND RESULTS

9.1.1 Conceptualization of Discourse-level Relations

In this work we conceptualize a specific type of discourse-level relation: relationships that exist between opinions by virtue of their related targets, and based on whether and in what way the opinions reinforce an overall stance. We create a linguistic representation that captures the variety of ways people express their opinions in a discourse. Our representation, the opinion frame, encodes information such as opinions of different polarities being employed to support the same overall stance. Our scheme formalizes the idea of alternative relations between targets and relations between opinions.

9.1.2 Annotation

The manual annotation studies test if our conceptualization of discourse-level relations can be recognized by humans. Overall, the results of our annotation experiments show that humans can label opinion types, opinion polarities and target links reliably. Annotators can retrieve opinion spans and target spans as well as in similar experiments carried out on text data by previous researchers. Likewise, the reliability of the task of target link detection is similar to that of coreference chain identification. Our annotation experiments help to identify the elements of the scheme that are clear and the parts that are challenging. We discover that determining whether items in the discourse are related is difficult and depends on the complexity of the discourse. Our results also indicate that annotation of opinion polarities can be very reliable when the targets are known.

Another contribution that is a by-product of our annotation endeavor is the implementation of the MPQA annotations on a multi-party face to face meeting corpus. Our annotation efforts confirm that the subjectivity scheme that was originally designed on news texts can be adapted and applied reliably to multiparty conversations. We find that the definitions and scheme pertaining to the sentiment category are applicable as is, but that arguing is manifested differently. We create extensions to the definition and annotation scheme for this category.

9.1.3 Features for Opinion Recognition in Meetings

We explore linguistic clues that can help the recognition of opinions in meetings. Our results validate that the sentiment lexicon created for monologic texts and our new arguing lexicon are useful for finding opinions. The results also suggest that dialog-level intention, conveyed by Dialog Acts, are useful indicators of opinions in meetings.

9.1.4 Fine-grained Polarity Disambiguation using Discourse-level information

This work is one of the first empirical studies to measure the impact of discourse-level relations on fine-grained polarity disambiguation. We measure the impact that perfect discourse-

based information can have on expression-level polarity disambiguation. Our results show that by augmenting word-based information with reliable discourse-level relation information, the polarity of difficult and ambiguous cases can be resolved effectively.

9.1.5 Computational Modeling of Discourse-level Relations for Fine-grained Polarity Disambiguation

This thesis explores ways in which a discourse-level scheme can be computationally modeled to produce a global inference. We explore a supervised collective classification framework and an unsupervised optimization framework for incorporating the discourse-level information for polarity classification. We find that both frameworks can effectively model the discourse-level relations to produce significant improvements over the local method. The success of the unsupervised optimization framework attests that the human insights behind our discourse relations provide coherence constraints that can be encoded effectively. The good performance of the supervised collective classification signifies that, by employing reliable discourse-level information, interdependent and discourse-coherent polarity interpretations can be learnt from annotations.

9.1.6 Features for Recognizing Discourse-level Relations

We create linguistic features based on discourse continuity, coreference, discourse relations, and dialog-level relations to identify and classify target relations and opinion-frame relations. Our results indicate that these features enable a simple supervised classifier to achieve better performance than baseline. However, we discover that the improvements are not sufficient for bootstrapping a global classifier. Due to this, automatically produced discourse-level relations are not able to produce improvements beyond the local methods for polarity classification.

9.1.7 Unsupervised Learning of Discourse-level Relations

We explore whether the ideas behind the discourse-level relations can be incorporated in an unsupervised method for stance classification of product debates on the web. This is achieved by using rules and web mining. We design rules for creating opinion target pairs and handling concessionary opinions. We find that our method that employs Point-wise Mutual Information learns the *same* target relations. We observe that our web mining approach effectively learns reinforcing relations, and also implicitly captures the *same* and *alternative* relations. This suggests that elements of our discourse-level relations can be learnt in an unsupervised fashion for certain domains.

9.1.8 Stance Classification in Dual-sided Debates

We attempt to classify stances in dual-sided debates using opinions and ideas from our discourse-level relations. We observe manifestations of target relations, as well as reinforcing and non-reinforcing relations in both product debates and political and ideological debates. In product debates, we are able to construct unsupervised stance classifiers that mine and employ these relations. Our systems perform better than a high-precision baseline approach.

However, we find that the abstract nature of our topics in the political and ideological domains, coupled with variability in beliefs makes stance classification challenging for political debates. Our supervised stance classifier using sentiment and arguing opinion-target features performs better than a distribution-based baseline, but it does not significantly outperform the current state of the art. Nevertheless, we observe that our systems are able to capture more insightful information regarding stance-taking than a system using unigrams. This lays the groundwork for future explorations.

9.1.9 Arguing Opinions

Arguing is a less well explored category in opinion analysis and, while this category is not the central focus of this thesis, as a byproduct, we gain a deeper understanding of this category. As a part of our work, we investigate how arguing is expressed in meetings, manually annotate

this category, and explore linguistic features that are useful for detecting arguing. We also investigate how arguing is expressed in political and ideological debates on the web and create a lexical resource for it. We find that arguing is a more prominent opinion category in such debates, and the system employing arguing features does better than the one employing sentiment features alone.

9.2 FUTURE DIRECTIONS AND OPEN PROBLEMS

9.2.1 Linguistic Scheme and Annotations

This thesis lays the groundwork for discourse-level relations between opinions. The insights gained from this exercise provide a number of directions for future explorations.

9.2.1.1 Target relations We define two types of target relations in this work: *same* and *alternative*. These relations are prolific in our meeting data as well as our dual-sided debates. However, it is possible for two items to be related but not have a strictly *same* or *alternative* relation. It is an open question what these relationships are and whether such relations are useful for opinion analysis.

Future work should also explore creating finer distinctions within the *same* target relation category. The *same* relation covers a number of linguistic phenomena such as coreference, synonymy, is-A relations, generalization, specialization, class-instantiation, epithets and other forms of bridging descriptions. Annotating this detail for the *same* relations will be useful from both a linguistic viewpoint as well as for applications. We conjecture that, apart from the linguistic phenomena listed above, more complex relations, such as *deictic* (observed in example 1.2), *cause-effect* and *requirement-enablement* will emerge once the clear cases of coreference and synonymy are accounted for. Also, explicitly annotating the different categories will provide insights into the types of features that can be employed for learning the target relations.

Fine-grained distinctions will also be useful if applications need to draw inferences re-

garding new relationships. We can explain this idea as follows: Consider three targets t_1, t_2 and t_3, where t_1 and t_2, and t_2 and t_3 are in the *same* target relation. Now, if both these relations are due to synonymy, an application can infer that t_1 and t_3 are also in the *same* (synonymy) relation. That is, the *same* target relations due to synonymy can be considered to be transitive. Now if a fourth target t_4 is *alternative* to t_2, it can be inferred to be *alternative* to t_1 and t_3 too. In a different scenario the inference will be different.

For example, for the same high-level target relations, if t_2 and t_3 are instantiations of t_1 (say for example t_1 denotes car, t_2 denotes Lexus, and t_3 stands for BMW), it is a fallacy in most situations to consider t_1 and t_3 as same; in fact it is quite likely that the discourse will set them up as alternatives. Now if a fourth target, t_4 (say Audi), is known to be an alternative to t_2, no inference should be drawn between t_4 and the remaining targets unless there is explicit evidence in the discourse for it. On the other hand an alternative to t_1 (say t_5, standing for train) can be inferred to be an alternative to t_2 as well as t_3. Thus, in this case, the inference can pass downwards (to the instantiations), but not necessarily between siblings.

9.2.1.2 Opinion relations In this work, the linguistic scheme includes the binary distinction between reinforcing and non-reinforcing relations. Later on, in our work on debates, we discovered that concessionary relations, which are a type of non-reinforcing relations, are particularly interesting for debate stances. It is an open question if other subcategories within the reinforcing and non-reinforcing categories have such impact. We conjecture this will depend on the application. For example, when two opinions are related by a reinforcing relation, and one is specifically a reasoning or justification for the other, a summarization system might prefer to pick the latter opinion. Future work can explore the finer distinctions between the opinion categories and their impact on applications.

9.2.1.3 Opinion relations independent of target relations In this work, the linguistic scheme considers relations between opinions when they have a target relation. As a first step, this policy ensures that the opinion frames capture the clear and tangible relations between the opinion expressions. However, opinions can be related even if they do not

exhibit an apparent target relation. This can happen in a number of scenarios, as shown in Example 3.8, where one of the opinions does not have a target and in Example 3.9, where the target relations do not fall under the *same* or *alternative* category. By removing the reliance on target relations, such cases, which are currently omitted, will be captured.

9.2.1.4 More consistent annotations for machine learning features In Chapter 6, we encountered difficulties in recognizing target relations. This is partly due to the fact that the input to the machine learning algorithm is noisy. This situation arises because target annotations are created with respect to the opinion annotations and target relations are created between entities that are targets and not otherwise. While this is sound from the perspective of opinion annotations, this approach results in a corpus where the same entities are inconsistently linked. This produces noisy instances for machine learning that attempts to use coreference features. In order to remedy this, we suggest an additional annotation step after the original target relations are established. This step can create consistent links between the targets and expressions that are coreferrent or bridging descriptions. The additional annotations will also alleviate the problem of large data skew.

9.2.2 Integer Linear Programming for Fully Automated Polarity Classification

In Chapter 6, we attempted to create an interdependent classification framework for target link classification, frame link classification and opinion polarity classification using a collective classification framework (ICA). Also recall that, in Chapter 5, we found that the ILP formulation performs better than ICA for polarity classification using reliable discourse information.

One of our future directions will explore an ILP formulation for the fully automated polarity classification. It is an open question whether an ILP framework will produce similar improvements when automatically generated discourse-level relations are used. It is possible that, due to the hard coding of discourse constraints, ILP's performance will get worse than ICA when the information is noisy. On the other hand, ILP can be designed to work on the probability distribution of the local classifiers (as against ICA, which uses only the

classifications of the local classifiers). Thus, weak (and hence presumably noisy) estimates from the local classifiers can be ignored or given less weight. In the ILP formulation using manual discourse information, the penalty for violating discourse constraints was set to a constant (1). Under the fully automatic setting, the weights for the different slacks in ILP can be changed to reflect the reliability of the classifiers.

9.2.3 Explicit Discourse-level Relations for Political and Ideological Debates

In this thesis, we perform supervised stance classification in political and ideological debates. In the supervised formulation, the classifier uses opinion target pairs as features and determines how to use these to recognize stances. Admittedly, this approach is limited – we do not find relations between opinion target pairs, only between opinion-targets and the stances. It will be interesting to see if reinforcing associations, similar to those in Chapter 7, can be learnt for these kinds of debates. However, due to the complexities of such domains, there are a number of challenges that will have to be handled.

Primarily, a more precise method of opinion target pairing is needed. The current approach of pairing the sentence-level opinion to all the content words in the sentence will produce a lot of noisy associations in web mining, especially if the sentences are long. The opinion target pairing can be limited to either clauses, or narrowed down by creating more detailed syntactic dependency rules.

Our experiments in this thesis reveal that arguing is prominently employed for expressing ideological stances. Improving the arguing lexicon will thus be important. One approach for this might be to adapt Josef Ruppenhoffer's lexicon from our previous work [Somasundaran et al., 2007a] to add to the existing arguing clues. This lexicon has regular expressions in addition to multi-word expressions. Some of the categories from this lexicon, such as necessity and emphasis, seem intuitive for arguing in online debates.

Finally, the approach to web mining will also have to be modified. Notice that when we analyzed the top features for Gay Rights debates in Section 8.6, we did not find opinions toward the main topic (for example, features such as *an-gay* or gay^{-}). Presumably, during web-mining, we will not find as many occurrences of explicit opinions towards the main

topics as we did for product debates. Additionally, as the stances are dependent on personal beliefs, the opinion-target pairs found in the debates may be too specific or uncommon for the domain. Thus, the web-mining process will need to cast a wider net if we hope to capture the idiosyncratic beliefs. We conjecture that in such situations, we will need to find chains of associations, rather than direct associations between opinion-targets and the opinion-topic. In order to get these chains, web-mining will have to be an iterative process. For example, a participant may argue against gay rights by arguing positively about corinthians. However, this might be a very rare occurrence. On the other hand, positive arguing regarding christianity and negative arguing regarding gay rights may be more popular. Web-mining may reveal a direct association between *an-gay* and *ap-christianity*, but not between *an-gay* and *ap-corinthians*. Multiple iterations of web mining can be employed to find that *ap-christianity* and *ap-corinthians* are associated. Consequently, *ap-corinthians* would be indirectly associated with *an-gay*. In this process the opinion target pairs whose affiliations are resolved become seeds for the successive stages of web mining. Iterative web mining can thus be employed to find reinforcement chains.

9.2.4 Multi-sided Debate Stance Classification

We have investigated stance classification in dual-sided debates. These debates are scenarios where our binary distinctions regarding target and opinion relations come to the fore. In the future, it will be interesting to investigate if these relations hold equally well in multi-sided debates. When there are multiple sides, each choice has multiple alternatives. This increases the number of ways to argue for a side and also adds more complexity.

For example, consider a three-topic three-sided debate. The topics are A, B and C and the corresponding sides are pro-A, pro-B and pro-C. A pro-A stance can be supported either by positive opinions regarding A, or by negative opinions regarding B or C; that is, by the opinion topic pairs: A^+, B^-, or C^-. Similarly, a pro-B stance can be supported by B^+, A^- or C^- and a pro-C stance can be supported by C^+, A^- or B^-. Thus, there are 6 opinion-topic pairs towards which, the associations of the opinion-targets (found in the posts) have to be mined. Notice that negative opinions towards a topic, say B^-, can indicate a support

	Positive	Negative
Accepted Items	120	20
Rejected Items	9	12

Table 42: Opinion Polarity Distribution for Accepted/Rejected Items

for either one of the two stances: pro-A and pro-C. This introduces additional ambiguity that will have to be resolved.

It is an open question if new relations or finer distinctions within the current relations will emerge from the analysis of multi-sided debates. For example, it is possible that ambivalence relations might become more prominent when there are multiple sides, as multi-sided debates might not polarize its participants as much as dual-sided debates.

9.2.5 Meeting Decision Prediction

Future directions can also explore how to employ our discourse-level relations in applications. Opinions and their discourse-level relations can help a decision prediction application for AMI meetings. This is because, in goal oriented meetings, participants discuss their viewpoints, present their sentiments, and argue with each other in an effort to reach a decision. Their opinion stances would thus shape the final decisions.

As a pilot study, we counted the number of positive and negative opinions towards the items that were accepted or rejected in the meetings. We obtained the information about accepted and rejected items, or meeting decisions, from the manual abstractive summaries provided by the AMI corpus. Then, we counted the opinions towards each of these items in the meetings (we used the manual annotations for this purpose). The items in the AMI meetings are mainly options for the new TV remote, which include attributes and features like different shapes, materials, designs, and functionalities.

Table 42 shows a contingency table of counts of positive/negative opinions for accepted/rejected items for 5 AMI meetings. We observe that the number of positive opinions

are more for accepted items, while the number of negative opinions are more for rejected items. Thus, positive and negative opinions towards items may be indicative of the decision taken regarding them.

The discourse-level relations provide more information regarding the stances. For instance, in Example 1.3, the speaker reinforces his stance supporting curved shapes for the remote via negative opinions towards the square shapes. The discourse-level relations bring in more opinions that are related to the stance. In fact, a meeting can be considered as a meta-document with stances towards various aspects of the remote control. These stances can influence the final decisions (accept/reject) regarding the aspects for the remote prototype.

BIBLIOGRAPHY

[Adamic and Glance, 2005] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. In *LinkKDD*.

[Agrawal et al., 2003] Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. In *WWW*.

[Akkaya et al., 2009] Akkaya, C., Wiebe, J., and Mihalcea, R. (2009). Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190–199, Singapore. Association for Computational Linguistics.

[Andreevskaia and Bergler, 2006] Andreevskaia, A. and Bergler, S. (2006). Semantic tag extraction from WordNet glosses. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-2006)*, Genova, Italy.

[Asher et al., 2008] Asher, N., Benamara, F., and Mathieu, Y. Y. (2008). Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters and Demonstrations*, pages 5–8, Manchester, UK. Coling 2008 Organizing Committee.

[Bansal et al., 2008] Bansal, M., Cardie, C., and Lee, L. (2008). The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*.

[Bethard et al., 2004] Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In Shanahan, J. G., Wiebe, J., and Qu, Y., editors, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US.

[Biber, 1988] Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

[Bilgic et al., 2007] Bilgic, M., Namata, G. M., and Getoor, L. (2007). Combining collective classification and link prediction. In *Workshop on Mining Graphs and Complex Structures at the IEEE International Conference on Data Mining*.

[Bloom et al., 2007] Bloom, K., Garg, N., and Argamon, S. (2007). Extracting appraisal expressions. In *HLT-NAACL 2007*, pages 308–315, Rochester, NY.

[Burger et al., 2002] Burger, S., MacLaren, V., and Yu, H. (2002). The isl meeting corpus: The impact of meeting type on speech style. Denver, CO. ICSLP-2002.

[Burger and Sloane, 2004] Burger, S. and Sloane, Z. A. (2004). The isl meeting corpus: Categorical features of communicative group interactions. In *NIST Meeting Recognition Workshop 2004*, Montreal, Canada. NIST.

[Carletta, 1996] Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

[Carletta et al., 2005] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*. AMI-108.

[Carvalho and Cohen, 2005] Carvalho, V. and Cohen, W. W. (2005). On the collective classification of email speech acts. In *SIGIR*.

[Choi et al., 2006] Choi, Y., Breck, E., and Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney, Australia. Association for Computational Linguistics.

[Choi and Cardie, 2009] Choi, Y. and Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 590–598, Singapore. Association for Computational Linguistics.

[Clark, 1975] Clark, H. H. (1975). Bridging. *Theoretical issues in natural language processing. New York: ACM.*

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Meas*, pages 37–46.

[Cohen, 1968] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–20.

[Cohen, 1987] Cohen, R. (1987). Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13:11–24.

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 168–175, Philadelphia, Pennsylvania.

[Dancygier, 2006] Dancygier, B. (2006). *Conditionals and Prediction: Time, Knowledge and Causation in Conditional Constructions*. Cambridge University Press.

[Denis and Baldridge, 2007] Denis, P. and Baldridge, J. (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL 2007*.

[Devillers et al., 2005] Devillers, L., Abrilian, S., and Martin, J.-C. (2005). Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In *Proc. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005)*.

[Devitt and Ahmad, 2007] Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic. Association for Computational Linguistics.

[Dhillon et al., 2003] Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. (2003). Meeting recorder project: Dialog act labeling guide. Technical report, ICSI Technical Report TR-04-002, Version 3.

[Ducrot, 1973] Ducrot, O. (1973). *Le preuve et le dire*. Mame.

[Efron, 2004] Efron, M. (2004). Cultural orientation: Classifying subjective documents by cocitation analysis. In *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*.

[Esuli and Sebastiani, 2005] Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM-05)*, pages 617–624, Bremen, Germany.

[Flammia and Zue, 1997] Flammia, G. and Zue, V. (1997). Learning the structure of mixed initiative dialogues using a corpus of annotated conversations. In *Eurospeech*, pages 1871–1874, Rhodes, Greece.

[Galley et al., 2004] Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *ACL*.

[Gamon and Aue, 2005] Gamon, M. and Aue, A. (2005). Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, US.

[Ganapathibhotla and Liu, 2008] Ganapathibhotla, M. and Liu, B. (2008). Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on*

Computational Linguistics (Coling 2008), pages 241–248, Manchester, UK. Coling 2008 Organizing Committee.

[Germesin and Wilson, 2009] Germesin, S. and Wilson, T. (2009). Agreement detection in multiparty conversation. In *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*, pages 7–14, New York, NY, USA. ACM.

[Goldberg and Zhu, 2006] Goldberg, A. B. and Zhu, X. (2006). Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*, New York, NY.

[Grefenstette et al., 2004] Grefenstette, G., Qu, Y., Shanahan, J. G., and Evans, D. A. (2004). Coupling niche browsers and affect analysis for an opinion mining application. In *Proceeding of RIAO-04*, Avignon, FR.

[Grosz and Sidner, 1986] Grosz, B. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3).

[Gyamfi et al., 2009] Gyamfi, Y., Wiebe, J., Mihalcea, R., and Akkaya, C. (2009). Integrating knowledge for subjectivity sense labeling. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2009)*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

[Haghighi et al., 2005] Haghighi, A., Toutanova, K., and Manning, C. (2005). A joint model for semantic role labeling. In *CoNLL*.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. In *SIGKDD Explorations, Volume 11, Issue 1*.

[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. pages 174–181, Madrid, Spain.

[Hillard et al., 2003] Hillard, D., Ostendorf, M., and Shriberg, E. (2003). Detection of agreement versus disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT-NAACL*.

[Hobbs, 1979] Hobbs, J. (1979). Coherence and coreference. *Cognitive Science 3(1)*, pages 67–90.

[Hobbs et al., 1993] Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1993). Interpretation as abduction. *AI*, 63.

[Hsueh and Moore, 2007] Hsueh, P.-Y. and Moore, J. (2007). What decisions have you made: Automatic decision detection in conversational speech. In *Proceedings of NAACL HLT 2007*, pages 25–32. Association for Computational Linguistics.

[Hu and Liu, 2004a] Hu, M. and Liu, B. (2004a). Mining and summarizing customer reviews. pages 168–177, Seattle, Washington.

[Hu and Liu, 2004b] Hu, M. and Liu, B. (2004b). Mining opinion features in customer reviews.

[Hu and Liu, 2006] Hu, M. and Liu, B. (2006). Opinion feature extraction using class sequential rules. In *Proceedings of AAAI 2006 Spring Sympoia on Computational Approaches to Analyzing Weblogs (AAAI-CAAW 2006).*

[Ikeda et al., 2008] Ikeda, D., Takamura, H., Ratinov, L.-A., and Okumura, M. (2008). Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP).*

[Joachims, 2005] Joachims, T. (2005). A support vector method for multivariate performance measures. In *ICML 2005.*

[Jurafsky et al., 1997] Jurafsky, D., Shriberg, E., and Biasca, D. (1997). *Switchboard-DAMSL Labeling Project Coder's Manual. http://stripe.colorado.edu/ jurafsky/ manual.august1.*

[Kamps et al., 2004] Kamps, J., Mokken, R. J., Marx, M., and de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, volume IV, pages 1115–1118, Paris, France. European Language Resources Association.

[Kanayama and Nasukawa, 2006] Kanayama, H. and Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363, Sydney, Australia. Association for Computational Linguistics.

[Kennedy and Inkpen, 2005] Kennedy, A. and Inkpen, D. (2005). Sentiment classification of movie and product reviews using contextual valence shifters. In *Proceedings of FINEXIN-05: Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations.*

[Kim and Hovy, 2004] Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. pages 1267–1373, Geneva, Switzerland.

[Kim and Hovy, 2007] Kim, S.-M. and Hovy, E. (2007). Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064.

[Kobayashi et al., 2005] Kobayashi, N., Iida, R., Inui, K., and Matsumoto, Y. (2005). Opinion extraction using a learning-based anaphora resolution technique. In *Proceedings of the*

2nd International Joint Conference on Natural Language Processing (IJCNLP-05), poster, pages 175–180.

[Krippendorff, 2004] Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology, 2nd Edition*. Sage Publications, Thousand Oaks, California.

[Kugatsu Sadamitsu and Yamamoto, 2008] Kugatsu Sadamitsu, S. S. and Yamamoto, M. (2008). Sentiment analysis based on probabilistic models using inter-sentence information. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

[Kunz and Rittel, 1970] Kunz, W. and Rittel, H. W. J. (1970). Issues as elements of information systems. In Univ. Stuttgart, I. F. G. d. P., editor, *Working Paper WP-131*.

[Landis and Koch, 1977] Landis, R. and Koch, G. (1977). The measurement of observer agreement for categorical data. In *Biometrics, Vol. 33, No. 1*.

[Laver et al., 2003] Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

[Lin, 2006] Lin, W.-H. (2006). Identifying perspectives at the document and sentence levels using statistical models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Doctoral Consortium*, pages 227–230, New York City, USA. Association for Computational Linguistics.

[Lin and Hauptmann, 2006] Lin, W.-H. and Hauptmann, A. (2006). Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1057–1064, Sydney, Australia. Association for Computational Linguistics.

[Lin et al., 2006] Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. (2006). Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-2006)*, pages 109–116, New York, New York.

[Liscombe et al., 2003] Liscombe, J., Venditti, J., and Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Eurospeech*.

[Litman and Forbes-Riley, 2006] Litman, D. J. and Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

[Lu and Getoor, 2003] Lu, Q. and Getoor, L. (2003). Link-based classification. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[Malouf and Mullen, 2008] Malouf, R. and Mullen, T. (2008). Taking sides: Graph-based user classification for informal online political discourse. *Internet Research*, 18(2).

[Mann and Thompson, 1988] Mann, W. C. and Thompson, S. A. (1988). *Rhetorical structure theory: Toward a functional theory of text organization*, chapter Text, 8 (3).

[Martin and White, 2005] Martin, J. R. and White, P. R. (2005). *The Language of Evaluation: The Appraisal Framework*. Palgrave Macmillan, London.

[Martin and Vanberg, 2008] Martin, L. W. and Vanberg, G. (2008). A robust transformation procedure for interpreting political text. *Political Analysis*, 16(1):93–100.

[McDonald et al., 2007] McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 432–439, Prague, Czech Republic. Association for Computational Linguistics.

[McDowell et al., 2009] McDowell, L. K., Gupta, K. M., and Aha, D. W. (2009). Cautious collective classification. *The Journal of Machine Learning Research*, 10:2777–2836.

[Mei et al., 2007] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07*. ACM.

[Miller, 1990] Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

[Miltsakaki et al., 2004] Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The penn discourse treebank. In *Language Resources and Evaluation Conference. Lisbon, Portugal. 2004.*

[Mohammad et al., 2009] Mohammad, S., Dunne, C., and Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 599–608, Singapore. Association for Computational Linguistics.

[Moschitti, 2009] Moschitti, A. (2009). Syntactic and semantic kernels for short text pair categorization. In *EACL*.

[Moschitti et al., 2006] Moschitti, A., Pighin, D., and Basili, R. (2006). Semantic role labeling via tree kernel joint inference. In *CoNLL*.

[Mueller and Strube, 2001] Mueller, C. and Strube, M. (2001). Annotating anaphoric and bridging relations with mmax. In *2nd SIGdial Workshop on Discourse and Dialogue*.

[Mullen and Malouf, 2006] Mullen, T. and Malouf, R. (2006). A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*.

[Nakatani and Traum, 1998] Nakatani, C. and Traum, D. (1998). *Draft: Discourse Structure Coding Manual version 2/27/98*.

[Neiberg et al., 2006] Neiberg, D., Elenius, K., and Laskowski, K. (2006). Emotion recognition in spontaneous speech using gmms. In *INTERSPEECH 2006 ICSLP*.

[Neville and Jensen, 2000] Neville, J. and Jensen, D. (2000). Iterative classification in relational data. In *In Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20. AAAI Press.

[Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. pages 271–278, Barcelona, ES. Association for Computational Linguistics.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Vol. 2(1-2):pp. 1–135.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86, Philadelphia, Pennsylvania.

[Passonneau, 2004] Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *LREC*.

[Polanyi and Zaenen, 2005] Polanyi, L. and Zaenen, A. (2005). Contextual valence shifters. In *Computing Attitude and Affect in Text*. Springer.

[Polanyi and Zaenen, 2006] Polanyi, L. and Zaenen, A. (2006). *Contextual Valence Shifters*. Computing Attitude and Affect in Text: Theory and Applications.

[Popescu and Etzioni, 2005] Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 339–346, Vancouver, Canada.

[Popescu et al., 2005] Popescu, A.-M., Nguyen, B., and Etzioni, O. (2005). OPINE: Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 32–33, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

[Prasad et al., 2007] Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). *PDTB 2.0 Annotation Manual*.

[Purver et al., 2006] Purver, M., Ehlen, P., and Niekrasz, J. (2006). Detecting action items in multi-party meetings: Annotation and initial experiments. In *Machine Learning for Multimodal Interaction*, pages 200–211. Springer-Verlag.

[Raaijmakers et al., 2008] Raaijmakers, S., Truong, K., and Wilson, T. (2008). Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the 2008 Conference*

on Empirical Methods in Natural Language Processing, pages 466–474, Honolulu, Hawaii. Association for Computational Linguistics.

[Reidsma et al., 2006] Reidsma, D., Heylen, D., and Ordelman, R. (2006). Annotating emotions in meetings. In *Proc. of the fifth international conference on Language Resources and Evaluation, LREC 2006*, pages 1117–1122. ELRA.

[Richardson and Domingos, 2006] Richardson, M. and Domingos, P. (2006). Markov logic networks. *Mach. Learn.*, 62(1-2):107–136.

[Rienks et al., 2005] Rienks, R., Heylen, D., and van der Weijden, E. (2005). Argument diagramming of meeting conversations. In A., V. and J., O., editors, *Multimodal Multi-party Meeting Processing, Work-shop at the 7th International Conference on Multimodal Interfaces*, pages 85–092, Trento, Italy.

[Riloff and Wiebe, 2003] Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *EMNLP 2003*, pages 105–112, Sapporo, Japan.

[Roth and Yih, 2004] Roth, D. and Yih, W. (2004). A linear programming formulation for global inference in natural language tasks. In *Proceedings of CoNLL-2004*, pages 1–8. Boston, MA, USA.

[Sadamitsu et al., 2008] Sadamitsu, K., Sekine, S., and Yamamoto, M. (2008). Sentiment analysis based on probabilistic models using inter-sentence information. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[Sen et al., 2008] Sen, P., Namata, G. M., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. Technical report, UMD Technical Report CS-TR-4905.

[Snyder and Barzilay, 2007] Snyder, B. and Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *Proceedings of NAACL-2007*.

[Somasundaran et al., 2009a] Somasundaran, S., Namata, G., Getoor, L., and Wiebe, J. (2009a). Opinion graphs for polarity and discourse classification. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 66–74, Suntec, Singapore. Association for Computational Linguistics.

[Somasundaran et al., 2009b] Somasundaran, S., Namata, G., Wiebe, J., and Getoor, L. (2009b). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179, Singapore. Association for Computational Linguistics.

[Somasundaran et al., 2007a] Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2007a). Detecting arguing and sentiment in meetings. In *SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

[Somasundaran et al., 2008a] Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2008a). Discourse level opinion relations: An annotation study. In *SIGdial Workshop on Discourse and Dialogue*. ACL.

[Somasundaran and Wiebe, 2009] Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.

[Somasundaran and Wiebe, 2010] Somasundaran, S. and Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

[Somasundaran et al., 2006] Somasundaran, S., Wiebe, J., Hoffmann, P., and Litman, D. (2006). Manual annotation of opinion categories in meetings. In *ACL Workshop: Frontiers in Linguistically Annotated Corpora(Coling/ACL 2006)*, Sydney, Australia.

[Somasundaran et al., 2008b] Somasundaran, S., Wiebe, J., and Ruppenhofer, J. (2008b). Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 801–808, Manchester, UK.

[Somasundaran et al., 2007b] Somasundaran, S., Wilson, T., Wiebe, J., and Stoyanov, V. (2007b). Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *International Conference on Weblogs and Social Media*, Boulder, CO.

[Stent, 2000] Stent, A. (2000). Rhetorical structure in dialog. In *INLG '00: Proceedings of the first international conference on Natural language generation*, pages 247–252, Morristown, NJ, USA. Association for Computational Linguistics.

[Stone, 2000] Stone, P. (2000). *The General-Inquirer http://www.wjh.harvard.edu/~inquirer*.

[Stoyanov and Cardie, 2008a] Stoyanov, V. and Cardie, C. (2008a). Annotating topics of opinions. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[Stoyanov and Cardie, 2008b] Stoyanov, V. and Cardie, C. (2008b). Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 817–824, Manchester, UK. Coling 2008 Organizing Committee.

[Su and Markert, 2008] Su, F. and Markert, K. (2008). From word to sense: a case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, Manchester.

[Suzuki et al., 2006] Suzuki, Y., Takamura, H., and Okumura, M. (2006). Application of semi-supervised learning to evaluative expression classification. In *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006)*, pages 502–513, Mexico City, Mexico.

[Takamura et al., 2005] Takamura, H., Inui, T., and Okumura, M. (2005). Extracting semantic orientations of words using spin model. In *Proceedings of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, US. Association for Computational Linguistics.

[Takamura et al., 2007] Takamura, H., Inui, T., and Okumura, M. (2007). Extracting semantic orientations of phrases from dictionary. In *HLT-NAACL 2007*, pages 292–299, Rochester, NY.

[Taskar et al., 2004] Taskar, B., Wong, M., Abbeel, P., and Koller, D. (2004). Link prediction in relational data. In *Neural Information Processing Systems*.

[Thomas et al., 2006] Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

[Titov and McDonald, 2008] Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *International World Wide Web Conference (WWW)*.

[Toulmin, 1969] Toulmin, S. (1969). *The Uses of Argument*. Cambridge University Press.

[Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. pages 417–424, Philadelphia, Pennsylvania.

[Turney and Littman, 2003] Turney, P. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

[van Eemeren and Grootendorst, 2004] van Eemeren, F. H. and Grootendorst, R. (2004). *A systematic theory of argumentation*. Cambridge University Press.

[Vieira and Poesio, 2000] Vieira, R. and Poesio, M. (2000). An empirically based system for processing definite descriptions. *Comput. Linguist.*, 26(4).

[Wiebe, 1990] Wiebe, J. (1990). *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. PhD thesis, State University of New York at Buffalo.

[Wiebe, 1994] Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

[Wiebe, 2002] Wiebe, J. (2002). Instructions for annotating opinions in newspaper articles. Department of computer science technical report tr-02-101, University of Pittsburgh.

[Wiebe et al., 1999] Wiebe, J., Bruce, R., and O'Hara, T. (1999). Development and use of a gold standard data set for subjectivity classifications ann. pages 246–253, College Park, Maryland.

[Wiebe and Mihalcea, 2006] Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings of COLING-ACL 2006*.

[Wiebe et al., 2004] Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

[Wilson, 2007] Wilson, T. (2007). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of private states*. PhD thesis, Intelligent Systems Program, University of Pittsburgh.

[Wilson, 2008a] Wilson, T. (2008a). Annotating subjective content in meetings. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[Wilson, 2008b] Wilson, T. (2008b). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh.

[Wilson and Wiebe, 2003] Wilson, T. and Wiebe, J. (2003). Annotating opinions in the world press. In *Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pages 13–22, Sapporo, Japan.

[Wilson and Wiebe, 2005] Wilson, T. and Wiebe, J. (2005). Annotating attributions and private states. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

[Wilson et al., 2005a] Wilson, T., Wiebe, J., and Hoffmann, P. (2005a). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada.

[Wilson et al., 2005b] Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005*.

[Witten and Frank, 2002] Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *SIGMOD Rec.*, 31(1):76–77.

[Wrede and Shriberg, 2003] Wrede, B. and Shriberg, E. (2003). Spotting hotspots in meetings: Human judgments and prosodic cues. In *Eurospeech*, Geneva.

[Yi et al., 2003a] Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003a). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceeding of ICDM-03, the 3ird IEEE International Conference on Data Mining*, pages 427– 434, Melbourne, US. IEEE Computer Society.

[Yi et al., 2003b] Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003b). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003)*, pages 427–434, Melbourne, Florida.

[Yu and Hatzivassiloglou, 2003] Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP 2003*, pages 129–136, Sapporo, Japan.