6-2014

Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization

Adam Robert Faulkner
Graduate Center, City University of New York

AUTOMATED CLASSIFICATION OF ARGUMENT STANCE IN STUDENT ESSAYS: A LINGUISTICALLY MOTIVATED APPROACH WITH AN APPLICATION FOR SUPPORTING ARGUMENT SUMMARIZATION

by

ADAM ROBERT FAULKNER

A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2014

This manuscript has been read and accepted for the
Graduate Faculty in Linguistics in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy

Date

Dr. Martin Chodorow
Chair of Examining Committee

Date

Dr. Gita Martohardjono
Executive Officer

Dr. William McClure
Dr. Virginia Teller
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

AUTOMATED CLASSIFICATION OF ARGUMENT STANCE IN STUDENT ESSAYS: A LINGUISTICALLY MOTIVATED APPROACH WITH AN APPLICATION FOR SUPPORTING ARGUMENT SUMMARIZATION

by

Adam Robert Faulkner

Adviser: Professor Martin Chodorow

This study describes a set of document- and sentence-level classification models designed to automate the task of determining the argument stance (*for* or *against*) of a student argumentative essay and the task of identifying any arguments in the essay that provide reasons in support of that stance. A suggested application utilizing these models is presented which involves the automated extraction of a single-sentence summary of an argumentative essay. This summary sentence indicates the overall argument stance of the essay from which the sentence was extracted and provides a representative argument in support of that stance.

A novel set of document-level stance classification features motivated by linguistic research involving stancetaking language is described. Several document-level classification models incorporating these features are trained and tested on a corpus of student essays annotated for stance. These models achieve accuracies significantly above those of two baseline models. High-accuracy features used by these models include a dependency subtree feature incorporating information about the targets of any stancetaking language in the essay text and a feature capturing the semantic relationship between the essay prompt text and stancetaking language in the essay text.

We also describe the construction of a corpus of essay sentences annotated for supporting argument stance. The resulting corpus is used to train and test two sentence-level classification models. The first model is designed to classify a given sentence as a supporting argument or as not a supporting argument, while the second model is designed to classify a supporting argument as holding a *for* or *against* stance. Features motivated by influential linguistic analyses of the lexical, discourse, and rhetorical features of supporting arguments are used to build these two models, both of which achieve accuracies above their respective baseline models.

An application illustrating an interesting use-case for the models presented in this dissertation is described. This application incorporates all three classification models to extract a single sentence summarizing both the overall stance of a given text along with a convincing reason in support of that stance.

Acknowledgments

This dissertation has benefited from the support and encouragement of friends, family, colleagues, and teachers, all of whom helped to shape the ideas that eventually found their way into the final text.

I'd first like to acknowledge the kindness, patience, and invaluable guidance of Martin Chodorow. Martin steered this dissertation away from more than a few ill-conceived blind alleys and guided much of the statistical analysis. I'd also like to thank the other members of my committee, Bill McClure and Virginia Teller. Bill, in particular, helped to guide me through a rocky period in my graduate school career and has always displayed a willingness to delve into some of the more arcane aspects of NLP.

Many other members of the CUNY and Georgetown faculties helped to guide the course of my research interests over the years. Alan Hausman's class on Wittgenstein introduced me to mathematical logic and analytic philosophy and John Roy's syntax classes introduced me to theoretical linguistics. In grad school, Paul Portner, Bill McClure, Janet Fodor, William Sakas, and Martin Chodorow all helped me to discover formal and quantitative approaches to natural language modeling. In a way, this dissertation, which draws on linguistic analyses of student writing to guide the creation of automated classification features, was inspired by the inter-disciplinary spirit I discovered in their classes and seminars.

My two years with the Text, Language, and Computation group at Educational Testing Service spurred my initial interest in the computational modeling of student writing. Jill Burstein, Joel

Tetreault, Nitin Madnani, and Beata Beigman Klebanov all helped me to hone the programming and data analysis skills that made this dissertation possible. In particular, Jill Burstein was an unfailingly patient mentor.

Thanks are also due to the anonymous reviewers of FLAIRS 27, Applied Natural Language Processing track. Many of their insightful suggestions and comments were integrated into the final draft.

My family has been amazingly supportive throughout my graduate school career. My mother served as both stylistic critic and proofreader of nearly every major paper I've written over the past six years. My wife Becca has shown incredible patience each step of the way, as has my daughter Talia, who spent her second year of life competing with a dissertation for attention. My thanks and gratitude to all of you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This is a study of automated argument stance classification and supporting argument summarization. Using a set of linguistically motivated classification features, this study presents three high-accuracy classification models and a proof-of-concept supporting argument summarization system. There are a number of phrases in those two opening sentences that require explanation: *argument stance classification*, *supporting argument summarization*, and *linguistically motivated classification features*. This introductory chapter unpacks each of these phrases and, along the way, motivates our choice of topic and methodology.

1.1 Argument stance classification

Authors of persuasive texts such as reviews, editorials, and argumentative essays, have various goals in mind when articulating their view of a movie, book, policy change, or controversial claim. In the case of a movie or book review, the goal might be to convince the reader that the movie or book under discussion is or is not worth viewing or reading, while writers of editorials and argumentative essays are usually interested in convincing the reader of the truth or likelihood of claims such as *The death penalty should be abolished*. In the latter case, we say that these writers are articulating an *argument stance* either *for* or *against* a claim. Readers generally have little trouble in recognizing that a given text is written with the goal of convincing readers of the truth or likelihood of a claim. Further, when asked to infer the *for* or *against* stance of a piece of argumentative text, readers tend to agree much of the time (Wilson, 2008). In an automated classification setting, the goal is to determine if a computer can be trained to make these sorts of inferences. As with most tasks in Natural Language Processing (NLP), what is natural or obvious to a speaker or reader can only be modeled with great difficulty by a computer. This is particularly true in the case of argumentative text. Writers of this text variety make subtle use of a large inventory of linguistic resources, many of which have been described by researchers at the lexico-syntactic (Chafe and Nichols, 1986; Biber and Finegan, 1989; Martin and White, 2005), semantic (Martin and White, 2005, sec. 1.5), and discourse levels (Crismore and Farnsworth, 1989; Hyland, 2004, 2005; Du Bois, 2007). The first challenge for a classification model designed to recognize and classify the stance of argumentative text is to capture these linguistic subtleties as classification features.

Argument stance classification is a sub-task of a broader classification task in NLP, that of Sentiment Analysis, or the modeling of writer or speaker affect as it occurs in speech or text. The flood of Sentiment Analysis research over the past decade has been largely driven by commercial concerns: companies are interested in mining the massive store of user-generated text data found online for commercially exploitable information such as customer preferences and competitive intelligence. Even in its earliest phase, these systems achieved high-accuracy results using the most basic of feature sets (ngrams), and it has proven difficult to improve upon these early results by more than a few percentage points (Pang and Lee, 2008; Liu, 2012). In a sense, then, the initial problem of determining the opinion polarity of a document or sentence with high-accuracy has been solved. Researchers have moved on to open problems in Sentiment Analysis that deal with more subtle semantic information in text, such as the existence of "hedging" (Farkas et al., 2010), distinguishing factual from non-factual events (Saurí and Pustejovsky, 2009), and argument stance classification. In the case of argument stance classification, there have been only a handful of attempts at seriously grappling with this task (Somasundaran and Wiebe, 2010; Somasundaran, 2010; Anand et al., 2011; Walker et al., 2012a,b; Hasan and Ng, 2013a,b). Additionally, there exist few benchmark systems or descriptions of corpus and knowledge resource construction that could serve as potential guidelines for future researchers. The single example of a bona fide corpus of argumentative text is that of Walker et al. (2012b). However, this corpus consists of noisy internet forum data that can hardly be considered representative of argumentative text. The study herein tries to fill these lacunae in the research literature by presenting a complete account of the creation of a corpus of quintessentially argumentative text—student essays—annotated for document- and sentence-level stance. I also present a set of classification features that capture the unique linguistic features of argumentative language. This approach differs from other extant approaches in the literature, many of which attempt to port standard opinion mining features used in the review domain to the very different domain of argumentative text.

1.2 Supporting argument summarization

As readers, when presented with a writer's arguments regarding some claim, we are usually interested in more than simply the writer's *for* or *against* stance toward that claim. We are also interested in *why* the writer takes a particular stance toward the claim. What reasons does the writer give in support of their argument? These reasons are colloquially known as *supporting arguments*. Additionally, we might be interested in discovering the supporting arguments that are most representative of a text's overall argument. One real-world scenario involving this requirement is the current drive to discover the arguments and reasons supporting those arguments contained in "eRulemaking" data (Kwon et al., 2006; Cardie et al., 2008; Purpura et al., 2008). These data consist of troves of emails and forum comments arguing *for* or *against* various regulatory initiatives. Summary representations of the *for* and *against* reasons contained in these data could potentially allow rulewriters to evaluate public and stakeholder arguments in a more time-efficient manner.

In constructing a system that, given a student essay, returns a representative supporting argument for the overall stance of that essay, we take our cue from two strands of Sentiment Analysis research: *opinion reason* mining (Kim and Hovy, 2006) and *single-sentence summarization* of product reviews (Glaser and Schütze, 2012). The goal of *opinion reason mining* is to discover which sentences of a review give a reader the best sense of why the reviewer had a positive or negative opinion of some product. *Single-sentence summarization* systems provide an end-user with a single sentence extracted from the original review that indicates the overall opinion of the review (positive or negative) and also provides a good reason for this opinion. In an analogous way, the single-sentence summarization system presented in the second half of this study outputs a single supporting argument that indicates the writer's overall argument stance along with a representative supporting argument for that stance. As far as we are aware, this study constitutes the first attempt at constructing supervised classification models that automatically identify and classify the stance of supporting arguments in argumentative text. The use of these models in a single-sentence summarization system is also novel.

1.3 Linguistically motivated classification features

The relationship between Linguistics and NLP is one of the more contentious cross-disciplinary relationships in Artificial Intelligence (AI). While the relationship between formal logic and AI research in automated reasoning and planning has been more or less constant and fruitful over the past half century, the relationship between Linguistics and NLP, as summarized in Jurafsky and Martin (2009, pgs. 9-13), has been an on-again-off-again affair. From an early, collaborative phase in the 1950's and 1960's, characterized by "symbolic" or rule-based approaches to NLP tasks, to a more sectarian phase of mutual rejection after the empiricist, data-driven turn of the 1980's and 1990's, the emergence of NLP tasks involving the processing of linguistically subtle phenomena has led to a resurgence in collaboration between these two fields. This resurgence was first noted in Pereira (2000) and was subsequently explored in Uszkoreit (2009) and Moore (2009).

My study can be considered a contribution to the developing synergy between Linguistics and NLP. My approach is to motivate all classification features with linguistic research dealing with argumentative language. This approach is in marked contrast to much opinion-mining work which is less linguistically informed and relies largely on opinion-bearing word occurrence and ngram frequencies to determine the polarity of a piece of text. However, it is clear from even the briefest examination of argumentative text that argument stance is a more complex linguistic phenomenon than opinion and that traditional opinion-mining approaches may not be appropriate for this text variety. When dealing with opinion-bearing text, the author's positive or negative views can often be inferred by simply noting the percentage of positive (*good, great*) versus negative (*bad, terrible*) opinion-bearing words contained in the text. In contrast, inferring a writer's argument stance tends to be more complicated. As an example, consider the following statement regarding prison reform, taken from a set of essay prompts used by international students as part of an argumentative essay writing exercise.

(1) The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.

Asked to argue *for* or *against* the statements given in (1), a student's stylistic choices are partly constrained by the genre or situation-of-use conventions of the argumentative essay (Hyland, 1990). These conventions partially overlap with the more general organizational, discourse, and lexico-syntactic conventions of argumentative language (Hunston, 2010; Conrad and Biber, 2000). At the organizational level, a student might begin their essay with a thesis or position statement, as dictated by pedagogical instruction textbooks and style guides (Williams, 2007, pg. 195). At the lexico-syntactic level, the thesis statement itself generally contains the proposition being evaluated in the matrix clause (Hunston and Thompson, 2000, pg. 3), often containing material taken directly from the prompt text, along with an evaluative lexical item indicating the writer's position regarding that proposition. Both of these organizational and lexico-syntactic features are evident in the first sentence of a typical essay response to (1), given in (2).[1]

(2) The prison system is not outdated and I don't agree that society should not punish its criminals, but rehabilitate them.

Negation of the original proposition is a typical strategy for registering one's opposition to that proposition (Hunston and Thompson, 2000, pg. 15): *The prison system is outdated* becomes *The prison system is not outdated*. We can also infer the writer's opposition to the idea that prisoners should be rehabilitated rather than punished by the use of *I don't agree* with a clausal complement containing material from the second part of the prompt statement, *society should not punish its criminals, but [should] rehabilitate them*. Explicit self-attribution using *I* is the most direct way of positioning oneself (Du Bois, 2007, pg 143) in relation to the proposition being evaluated while use of the stance predicate *not agree* marks the complement as "disaligned" (Du Bois, 2007) with that proposition.

Reading further in the essay, we find one of the author's reasons for their opposition to the statements in the prompt, as given in (3).

(3) Criminals should be punished because if they won't be punished they won't stop to be criminals.

For the reader, making the inference that the sentence in (3) is a reason offered in support of the writer's opposition to the prompt rather than simply a restatement of that opposition involves two inter-related interpretive steps. The first step involves noting the location of (3) in relation to the thesis statement. Does (3) precede or follow that statement? The correct interpretation of (3) as a supporting argument for (2) is partly due to the reader's background knowledge regarding cohesive (Halliday and Hasan, 1976) inter-sentential structure: the canonical ordering in such contexts is thesis preceding supporting argument.

[1] Throughout this study, any grammatical or spelling errors contained in the original ICLE essays have been retained in examples extracted from those essays.

Second, correctly interpreting the dependent reason clause *if they won't be punished they won't stop be criminals* as *backing* (Toulmin, 2004) or *justification* (Mann and Thompson, 1988) for the rejection of *No civilized society should punish its criminals* involves noting the argument polarity (*for* or *against*) of the explicit statement of that rejection given in the thesis statement in (2), *I don't agree that society should not punish its criminals*. It is the reader's knowledge of the argument polarity of the thesis statement that allows them to make the inference that the proposition in the dependent clause is presented as a negative consequence of endorsing the prompt statement, and, in turn, is meant to provide support for the writer's rejection of that statement. Inter-sentential relationships of this kind are examples of the discourse-level resources utilized in argumentative writing.

The analysis we have just given of sentences (2) and (3) serves as a preview of the linguistic observations we use to motivate the classification features constructed in this study.

1.4 Contributions of this work

This study makes the following contributions:

- The construction of a corpus of student argumentative essays annotated for document-level argument stance polarity is described. This corpus is the first of its kind. Observed agreement and Cohen's κ score for this annotation task were .82 and .68, respectively.

- A high-accuracy document-level stance classification model is presented. Using a set of features motivated by influential linguistic accounts of argumentative language, our best-performing document-level stance classification model achieves an accuracy of 82%, which significantly improves upon two high baseline models.

- A novel annotation task is introduced: sentence-level annotation of supporting argument stance polarity. I describe the steps involved in tagging the stance polarity of supporting arguments contained in a a subsection of our stance-annotated student essay corpus. For this task, observed agreement and κ were .85 and .70, respectively.

- Two new classification tasks are introduced: identification of supporting argument sentences—a neutral vs. polar classification task—and classification of the argument stance polarity of supporting arguments. Models trained on linguistically motivated features are developed for both these tasks, each of which significantly improves upon high-accuracy baseline models. Our best-performing supporting argument identification model achieves an accuracy of 73% while our best supporting argument stance polarity classifier achieves an accuracy of 79%.

- The task of *supporting argument summarization* is introduced. This task involves extracting the single sentence of a given essay that indicates that essay's overall argument stance and also provides a good argument in support of that stance. A proof-of-concept system is presented that incorporates each of the document- and sentence-level classification models described in this study. The results of this system are evaluated using baseline sentences matched to each of the summary sentences. Crowsdouced raters preferred our system's summary sentence to the baseline sentence in 61.3% of cases, a result that was significant at level $p < .001$ (using the binomial test). Inter-rater observed agreement and Fleiss' κ score were .70 and .52, respectively,

1.5 Dissertation outline

This study is divided into three basic parts. In the first part, we describe how the NLP task of stance classification fits into the broader classification framework of Sentiment Analysis (chapter 2). We then provide an overview of influential linguistic accounts of stancetaking language (chapter 3) which we use to motivate a document-level stance annotation task (chapter 4) and a set of document-level stance classification features (chapter 5). Several document-level classification experiments using different combinations of these feature sets are then described (chapter 6).

The second part of this study targets another aspect of argumentative text: the writer's use of supporting arguments to buttress their *for* or *against* stance toward a given statement. We first provide a review of philosophical and linguistic research dealing with supporting arguments and show that supporting arguments in student argumentative text can be succinctly described at the lexico-syntactic and discourse levels using the frameworks provided in this research (chapter 7). We then motivate a new annotation task, supporting argument stance annotation, and a new classification task, supporting argument classification. The construction of a set of classification features for this task is described (chapter 8). These features are used to generate two distinct sentence-level classification models: a neutral-polar supporting argument classifier, which identifies any supporting arguments in a piece of argumentative text, and a supporting argument polarity classifier, which classifies a given supporting argument as providing a reason in support of a *for* or *against* stance (chapter 9).

In the final part of this study, we show that the classification models described in the first two parts can be incorporated into a single application: a supporting argument summarization system (chapter 10). Given a piece of argumentative text, this system is designed to extract a single supporting argument from the text that will give a reader a sense of the author's overall stance toward a statement and a good reason for that stance. We then conclude our study with some suggestions for future research (chapter 11).

Chapter 2

Sentiment analysis

2.1 Background

Sentiment Analysis is an NLP task concerned with the modeling of writer or speaker affect as it occurs in the language of reviews, essays, editorials, blogs, political speeches, and various other text varieties that deal with the subjective evaluation of an entity or proposition. One can point to two basic reasons for the NLP community's increased interest in this area over the past decade. First, the low-level tasks that preoccupied NLP researchers in the 1980's and 1990's, such as syntactic parsing and named entity recognition, have resulted in systems that were achieving accuracies well above 90% by the mid-2000's (McClosky et al., 2006; Ratinov and Roth, 2009). One consequence of these successes was a turn within NLP to unresolved semantic questions dealing with the automated identification of the writer's opinion (Pang and Lee, 2008), the factuality of a described event (Saurí and Pustejovsky, 2009), and argument stance (Anand et al., 2011; Somasundaran and Wiebe, 2010; Conrad et al., 2012; Walker et al., 2012b; Hasan and Ng, 2013a,b). In a sense, this shift constitutes a revival of interest in research concerns that had been neglected since Stone's (Stone et al., 1966) pioneering work on content analysis in the 1960's and Mosteller and Wallace's (1984) work on anonymous authorship identification in the 1980's.

A second reason for NLP's renewed interest in the computational modeling of evaluative language is the tremendous increase in computer-mediated commerce and social interaction we have witnessed since the beginning of the millenium. While this has led to an exponential increase in data, there has been comparatively little increase in information, where information is here understood as "interpreted data." In particular, the explosion of user-generated natural language data on the internet and the recognition that such data contain valuable information in the areas of business intelligence, strategic watch, product review analysis, and question-answer systems, has motivated a new set of questions for engineers and linguists working in NLP. Many of the questions addressed by these researchers are as old as Aristotle's *Rhetoric*: What aspects of language allow hearers and readers to infer the opinion or stance of a writer or speaker toward an entity or proposition? Which features of language allow speakers and writers to persuade their audience of the truth of an evaluative utterance? However, many questions are motivated by the immediate engineering concerns of commercial applications in the areas of review analysis and sentiment tracking: What features of evaluative language will allow a system to make the best predictions regarding a writer or speaker's attitude toward an entity or proposition? How can these features be generalized across domains as varied as movie reviews, editorials, survey responses, student essays, and blogs? All of these questions fall under the purview of Sentiment Analysis.

2.2 Defining sentiment

Although there is no universally agreed-upon definition of *sentiment* as it relates to automated text analysis, there is general agreement among linguists working within the framework of Systemic Functional Linguistics (Halliday and Matthiessen, 2004) regarding the existence of a register that allows speakers to express " 'private' states which can only be subjectively verified" (Quirk et al., 1985, pg. 202). This register most commonly occurs in contexts in which the speaker's interpretation of some phenomenon is unknown or is contested: reviews, editorials, and debates.

These text varieties all deal in “essentially contested concepts” (Gallie, 1955), such as “work of art” or “democracy,” that cannot be objectively validated as true or false. The language speakers use to validate such concepts tends to be emotion-laden or “sentiment-bearing.”

The first obstacle for any researcher attempting to define sentiment-bearing language is terminological confusion regarding this register. Like sentiment-bearing language itself, there are numerous, more or less interchangeable, terms used to describe the same phenomenon, each reflecting a particular linguistic focus. In Jespersen (1922), one finds a description of the role of grammatical mood that is indistinguishable from modern descriptions of sentiment-bearing language. Mood, for Jespersen, expresses “attitudes of the mind of the speaker towards the contents of the sentence” (Jespersen, 1922, pg. 313). The term *attitude* reappears in Halliday and Matthiessen (2004) where it is used to describe the role of the epithet in expressing a speaker’s “subjective attitude” toward an entity (Halliday and Matthiessen, 2004, pg. 318). Jakobson describes the emotive aspects of language as “a direct expression of the speaker’s attitude toward what he is speaking about” (Jakobson, 1960, pg. 4).

Labov’s (1972) discussion of narrative language stresses the speaker-oriented, or subjective quality of sentiment-bearing language. Subjectivity is encoded in narrative language using the devices of *internal evaluation*, which occurs while the story is being told (“I was shakin’ like a leaf”), and *external evaluation* which allows the speaker to suspend the narrative for a moment to report on their reaction to the event just described (“But it was quite an experience”). Lyons (1977) identified expressive language (“emotive” or “attitudinal” language) as the means by which “a speaker or writer establishes or reveals his individuality in a particularly original manner” (Lyons, 1977, pg. 107). Other commonly used terms for sentiment-bearing language include *affect* (Besnier, 1990), *opinion* (Pang and Lee, 2008; Liu, 2012), and *appraisal* (Martin and White, 2005). Identifying the linguistic features associated with a speaker’s implicit opinion, or *perspective* (Lin et al., 2006), can also be included under the rubric of sentiment-bearing language.

What all of these perspectives share is the general intuition that certain text varieties, such as reviews and editorials, allow writers to report their response to some phenomenon and, further, that this response is typically expressed using emotion-laden, or sentiment-bearing, language. Distinguishing the positive sentiment-bearing language of (4a,b) from the negative sentiment-bearing language of (4c,d) is largely a matter of recognizing a particular word's *prior polarity*, or its sentiment polarity (positive or negative) absent any sort of context (Wilson, 2008).

(4) a. Jack warden is **ideal** as Jehan, eliciting just the right amount of **humor** and **compassion** required for the character (the scenes involving him and Nello are **heartwarming**, **touching** ones). — From Pang et al. (2002)

b. Virtually **flawless**, with **impeccable** regional details, startlingly **original** characters, and a **compelling** literary plot that borders on the thriller, Ondaatje's **stunning** achievement is to produce an **indelible** novel of **dangerous beauty.** [USA Today Previews M Ondaatje *Anil's Ghost* Toronto: Vintage. 2000: i]
—From Martin and White (2005, pg. 43)

c. "Snake Eyes" is the most **aggravating** kind of movie: the kind that shows so much potential then becomes **unbelievably disappointing**. —From Pang et al. (2002)

d. This **abnegation** of the essence of democratic government goes much further than a **failure** to protect the nine British citizens who are incarcerated in this legal **black hole**. [Guardian, 10/01/04: leader pages 24]
—Martin and White (2005, pg. 168)

Reader inferences regarding the positively evaluated performance of actor Jack Warden and the quality of the novel *Anil's Ghost* in (4a) and (4b), respectively, are mainly due to positive-polarity lexis such as *compassion*, *heartwarming*, *flawless*, and *impeccable*, while the negative evaluations of the movie *Snake Eyes* and the policies of the British government in (4c) and (4d), respectively, turn on the largely negative lexis *aggravating*, *unbelievably disappointing*, *abnegation*, *failure*, *black hole*, etc. As is commonly the case, both (4a,b) and (4c,d) display a mix of positive and negative vocabulary, but most readers can easily make the correct inference regarding the polarity of the author's sentiment toward the films, books, acting performances, and policies under discussion. The goal of Sentiment Analysis is to automate such inferences.

2.3 Methods in Sentiment Analysis

Inferring the writer's opinion by simply noting sentiment-bearing term presence or frequency, as in (4), exemplifies the lexical focus that dominates current approaches to Sentiment Analysis. The earliest modern Sentiment Analysis research, such as Stone's General Inquirer system (Stone et al., 1966), made use of manually compiled lexicons of vocabulary annotated for positive and negative sentiment and this approach has several contemporary implementations (Wilson and Wiebe, 2005; Tong, 2001; Taboada et al., 2011). More commonly, however, sentiment lexicons are compiled automatically. Hatzivassiloglou and McKeown (1997) created a lexicon of sentiment-bearing adjectives by seeding a large corpus with adjectives manually identified as positive or negative. Depending on the pattern of conjunction observed in the corpus, that adjective's label is iteratively propagated to any conjoined adjectives (for example, *well-received* receives a positive label propagated from the positive-labeled *simple* since it appears in *simple and well-received*). A bootstrapped approach to attitude lexicon creation was also used in Kim & Hovy (2004) who began with a list of prototypically positive and negative adjectives, such as *good* and *bad*, and iteratively expanded this list using WordNet (Miller, 1995; Fellbaum, 1998) synonyms.

Sentiment Analysis systems dealing with genres containing highly domain-dependent vocabulary, such as the movie review corpus of Pang et al. (2002), have achieved accuracies above 80% (Pang et al., 2002; Mullen and Collier, 2004) using basic statistical models based on sentiment-bearing term presence or frequency. However, applying the same techniques to genres displaying more subtle expressions of writer sentiment introduces a new set of challenges, in particular, the challenge of polarity ambiguity at the domain and lexico-syntactic levels. Domain-dependent polarity ambiguity is exemplified by the use of *fascinate* in (5). While the typical use of *fascinate* is to convey positive sentiment, as in (5a), taken from a book review, *fascinate* can also be used negatively, as in (5b), taken from an editorial.

(5) a. At several different layers, it's a **fascinating** tale. —From Wiebe et al. (2004, pg. 279)

b. We stand in awe of the Woodstock generations ability to be unceasingly **fascinated** by the subject of itself. —From Wiebe et al. (2004, pg. 279)

Lexico-syntactic polarity-ambiguity problems also include phrase-level polarity shifts, as in (6a) where the negative-polarity *distortion problems* is shifted to positive, and the subtle use of negation in (6b), taken from newswire text, where the negated versions of *succeeded* and *succeed* maintain their positive polarities. Since *not succeeded* and *never succeed* are modified by the negative *breaking their will*, the negative-polarity act of breaking their will is predicted to fail, which, in turn, is viewed as something positive. Polarity shifts can also occur intersententially as occurs in the thwarted expectations narrative of (6c), a common rhetorical strategy in review text.

(6) a. I have **not** had any **distortion problems** with this phone and am more pleased with this phone than any I've used before. —From Ikeda et al. (2008, pg. 2)

b. They have **not succeeded**, and will **never succeed**, in breaking the will of this valiant people. —From Wiebe et al. (2005)

c. This movie should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a **good** performance. However, it can't hold up. —From Pang and Lee (2008, pg. 22)

To deal with examples such as (6) researchers have developed both rule-based (Choi and Cardie, 2009) and machine learning approaches (Ikeda et al., 2008) that handle polarity reversals resulting from multiply negated sentiment-bearing words. Wilson et al. (2009) presented a phrase-level opinion analysis system designed to deal with examples such as (6b). The feature set of Wilson et al. was built using newswire text annotated for contextual polarity effects. The highest resulting accuracy for various machine learning algorithms trained on this feature set was 74.5%. Systems designed to capture the discourse-level effects of (6c), where global (negative) sentiment is the result of sentence ordering rather than word frequency, include sequential models of intersentential “sentiment flow” (Mao and Lebanon, 2006), cohesion-based approaches (Devitt and Ahmad, 2007), and polarity classification systems that incorporate document-level (Taboada et al., 2008; Zirn et al., 2011; Trivedi and Eisenstein, 2013) and sentence-level (Heerschop et al., 2011) discourse relations.

2.4 Sub-tasks in Sentiment Analysis

2.4.1 Aspect-based Sentiment Analysis

Other, more fine-grained sub-tasks within Sentiment Analysis have emerged in recent years, such as “aspect-based” Sentiment Analysis, perspective classification, and stance classification. In the case of aspect-based Sentiment Analysis, the goal is to identify, not simply the global sentiment of a particular document or sentence, but the sentiment associated with aspects of the target of that global sentiment. A review of a phone, for example, contains an overall opinion polarity, but could also contain opinions regarding different aspects of the phone (reception, user-interface, or camera). Identifying and classifying aspect-based opinions involves first identifying aspect targets, which in turn involves problems of co-reference resolution and the induction of meronymic (part-of) hierarchies of the global opinion target (Pang and Lee, 2005; Goldberg and Zhu, 2006; Snyder and Barzilay, 2007). For instance, *it* in *it has the best touchscreen* should be correctly linked with *iPhone*, and *touchscreen* should be correctly identified as being part of the global opinion target, *iPhone*.

2.4.2 Perspective classification

Perspective classification, like opinion mining, is a two-class classification problem but the target classes in this case are socio-political perspective (e.g., conservative vs. liberal, Palestinian vs. Israeli) rather than positive or negative polarity. Malouf and Mullen (2008) analyzed online political debate data, self-tagged by writers as conservative or liberal. They identified the quoting relationship between posters as the most informative feature for political perspective classification. The point of quote-and-response is to rebut an opposing perspective and thus simply identifying the political perspective of the quote source in turn serves to identify the quoting poster as holding the opposite political perspective. Lin et al. (2006) experimented with a word-usage feature to identify Israeli versus Palestinian perspective in a corpus of documents collected from an online debate site. In this case, word-appearance and word-frequency proved informative. For example, the words *Palestinian* and *occupation* appear more frequently in documents written from the Palestinian perspective than in documents written from the Israeli perspective; on the other hand, the words *Israeli* and *secure* appear more frequently in documents written from the Israeli perspective. Perspective-specific keywords were also discovered to be highly discriminative in experiments automatically classifying perspectives on the death penalty (Klebanov et al., 2010).

2.4.3 Stance classification

Stance classification, which involves the automated classification of the writer's positive (=*for*) or negative (=*against*) stance toward a given proposition, is the most recent addition to the array of sub-tasks associated with Sentiment Analysis. The most significant work dealing with the automated classification of argument stance is that of Somasundaran and Wiebe (2010), Anand et al. (2011), Walker et al. (2012a), and Hasan and Ng (2013a). These researchers have introduced several supervised approaches to this task using a corpus of online debates. Although there are abundant annotated data for traditional opinion mining and Sentiment Analysis tasks, there are no comparable resources for stance classification work. Thus, we see in this work a reliance on data collected from popular online debate forums dealing with controversial issues such as gun control, abortion, the existence of God, and gay marriage. This data is freely available and abundant—the *Internet Arguments Corpus* of Walker et al. (2012b), for example, contains over 390,000 debate posts. Additionally, since users must tag their debate posts as *for* or *against*, these data are, in effect, self-annotated for argument stance. In these forums, users are invited to create debate topics that can be evaluated via the *for* or *against* distinction. This eliminates debate topics that cannot be evaluated using this distinction, such as "How do you think President Obama is doing?" or "When will we get to Mars?" Instead, topics must be framed as bare assertions, "Abortion should be legal," "God exists," or as yes/no questions, "Should abortion be legal?," "Does God exist?"

The system of Somasundaran and Wiebe (2010) presents a supervised approach to the classification of stance in these online debate data. A significant aspect of the feature set developed in Somasundaran and Wiebe is the use of a stance lexicon comparable to the opinion lexicons traditionally used in opinion mining tasks (see section 4.6). This lexicon was constructed using annotations from the Multi-Perspective Question Answering Project (MPQA) (Wiebe et al., 2005). Additionally, all debate text was pre-processed in such a way that information regarding the target of a given expression of argument stance (in *the death penalty should be abolished*, the target is *death penalty* and the expression of stance polarity is *should*) is directly appended to all stance-taking expressions in the text. Along with stance-target features, the system contains features recording information regarding the targets of opinion. Ablation experiments were performed using combinations of stance-target and opinion-target features represented as a frequency-valued bag-of-words vector. The highest accuracy of 64% was achieved by training an SVM learner on a combination of stance-target and opinion-target features.

The online debate exchanges collected by Walker et al. (2012b) typically take one of two forms, main topic response, in which a poster writes in direct response to the debate topic, or quote-and-response, in which posters quote all or part of the main topic response of a previous poster and attempt to rebut it. An example of the former is given as (7).

(7) a. Abortion should be banned.

b. Abortion shouldn't be banned. If abortion is considered by a woman, it is evidently due to the fact that the mother doesn't want the baby, or is facing a difficult time in their life such as those of teenage pregnancies etc. Without abortion, the child will be in jeopardy.

A corpus of debate forum posts containing such exchanges was collected by Walker et al. and served as a test bed for the classification system presented in Anand et al. (2011) and Walker et al. (2012a). The goal of these systems is to classify debate posts as argument or rebuttal rather than as positive (=*for*) or negative (=*against*) stance. Since rebuttals have a quote/response structure, one would expect to find greater usage of pronominalization, in particular frequent use of *you*, along with frequent use of propositional anaphora. The feature set utilized in this system included unigram and bigram counts, grammatical category features derived from the Linguistic Inquiry and Word Count toolkit (Pennebaker et al., 2001), syntactic dependencies capturing stance-target relationships inspired by similar work in Wilson et al. (2009), and generalized dependencies (Joshi and Penstein-Rosé, 2009). In the lattermost feature representation framework, one or both members of a dependency tuple, such as $\langle overwhelming, evidence\rangle$, is part-of-speech generalized to, for example, $\langle ADJ, evidence\rangle$. Accuracies were reported relative to topic domain with the highest accuracy of 69$\%$ achieved on 2nd Amendment debate data. Weka's (Hall et al., 2009) implementations of a Naive Bayes classifier and the RIPPER rule-learner (Cohen, 1995) were used in all experiments. Pronoun and anaphora counts along with positive and negative sentiment-bearing word counts were the most helpful discriminators of arguments versus rebuttals in their system.

Chapter 3

The language of stancetaking

Our focus in this study is sentiment-bearing language as it appears in student essays. In this chapter we will introduce the corpus used as a test bed for the experiments reported here, the International Corpus of Learner English (ICLE) (Granger, 2003). Using examples from this corpus, we will show that the language of argumentative essays is best described as *stancetaking* and so automating the process of determining the stance polarity of an essay should be considered a stance classification task along the lines of Somasundaran and Wiebe (2010) and Anand et al. (2011). We will also present a linguistic profile of stancetaking language as it appears in argumentative essays and we will argue that stancetaking simultaneously encodes expressions of opinion, which we will call *attitudinal* language, and stance. This will allow us to make the claim that stancetaking is the act of arguing *for* or *against* an attitudinal assessment. This conception of stancetaking motivates the creation of the classification features described in section 4.2.

3.1 Background

The question of how speakers and writers persuade their audience of the truth of a given claim dates at least to Aristotle, whose *Rhetoric* describes the "art of persuasion" used by orators to convince their audience of the validity of an argument. Aristotle's framework describes three rhetorical devices common to persuasive arguments: *ethos* (credibility of the speaker), *pathos* (emotional appeal), and *logos* (formal schemes of argumentation). These basic categories dominated the study of persuasive language until the modern period. Conceptual variants can still be found in pedagogical instruction textbooks and "good writing" manuals (Williams, 2007).

Several recent approaches to the computational modeling of evaluative language can be described using Aristotle's three rhetorical devices. Aristotle's *ethos* comes into play in work dealing with "epistemic stance" or the sourcing of evaluative utterances and with the automated classification of the factuality of an utterance (Saurí and Pustejovsky, 2009), which is assessed relative to the speaker of that utterance. *Pathos* (language meant to provoke emotional response) is roughly equivalent to the attitudinal form of evaluative language described above and remains the major focus of research in Sentiment Analysis and opinion mining. The relationship between stancetaking language and *logos*, or argumentation schemes, has been mostly ignored by researchers. Exceptions include recent NLP work involving the parsing of legal text (Reed et al., 2008; Palau and Moens, 2009) and news reportage and court documents (Feng and Hirst, 2011) using Walton's (2008) argumentation schemes.

3.2 Stancetaking in Systemic-Functional Linguistics

The idea that language has a stancetaking function has always been at the forefront of linguistic traditions stressing the social aspects of language use, such as Systemic-Functional Linguistics (SFL). Much current research on stancetaking language has occurred within the SFL framework. In itself, this is not surprising. If one makes the assumption, as SFL does, that meaning is non-truth-conditional and occurs as an "interactive event involving speaker, or writer, and audience" (Halliday and Matthiessen, 2004, pg. 106), the consequence is a focus on language use that most clearly involves a speaker or writer engaging a listener or reader who must be convinced of the writer's argumentative position. The majority of this research deals with transcribed speech corpora such as conversational exchanges (Conrad and Biber, 2000; Channell, 2000; Scheibman, 2007; Karkkainen, 2007), dialogue in movies (Martin, 2000), and academic lectures (Biber, 2006). This focus on speech is partly due to the fact that both sentiment-bearing and stancetaking language occur with observably higher frequency in speech than in written registers (Biber et al., 1999, pg. 979). Text corpora used in stancetaking language research dealing with written registers have included academic research articles (Hunston, 1989, 1993, 1994), issues of *New Scientist* magazine (Hunston, 2010), Darwin's *The Origin of Species* (Crismore and Farnsworth, 1989), school textbooks (Crismore, 1989; Hyland and Tse, 2004), company annual reports (Hyland, 1998), and newspaper reportage (Conrad and Biber, 2000; Thompson and Zhou, 2000).

3.3 Defining stancetaking

Many of the terminological problems we observed when discussing the concept of sentiment in section 2.1 reappear when we try to define stancetaking. Biber et al. (1999) define *stance* as the set of linguistic mechanisms that allow speakers and writers to express "personal feelings, attitudes, value judgments, or assessments" (Biber et al., 1999, pg. 966). Yet this definition is similar to Hunston and Thompson's (2000) definition of *evaluation* as "the broad cover term for the expression of the speaker or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about" (Hunston and Thompson, 2000, pg. 5). Biber et al. (1999) and Conrad and Biber (2000) divide stancetaking language into three basic sub-categories: epistemic stance, style stance, and attitudinal stance. Complicating matters, in Martin and White's (2005) influential taxonomy of the evaluative lexicon, Conrad and Biber's (2000) term *attitudinal* is used to describe the language of "emotion, ethics and aesthetics" (Martin and White, 2005, pg. 42). Researchers with a discourse, rather than lexical, focus view stance as an interactional phenomenon. In this view, to take a stance involves "simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects" (Du Bois, 2007, pg. 163). Aligning or disaligning with other speakers with regards to the object under discussion allows speakers to "negotiate their points of view and alignment with each other" (Keisanen, 2007, pg. 253). To capture the notion of stance as a discourse act, the term *stancetaking* is used.

A major reason that there are so many different terms for stancetaking language is that language encodes stance along many different dimensions. Emphasis on one dimension rather than another often leads researchers to conflate the features associated with that dimension with stancetaking language itself. Thus, the emphasis on single- and multi-word markers of stance (*certainly, the evidence suggests, I believe that, as X has demonstrated*) allows some researchers to claim that stancetaking language can be reduced to a single branch of their taxonomy of the English evaluative lexicon. In turn, the emphasis on discourse context—participants, alignment, etc. —leads researchers with a discourse focus to emphasize stance as an act among discourse participants rather than as a lexical or syntactic feature.

In this study, we take the view that stancetaking, as it appears in the language of argumentative essays, is simultaneously encoded across both discourse and lexico-syntactic dimensions. In the next few sections, we will develop this claim by first introducing the ICLE corpus and then describing the semantic, lexico-syntactic features, and discourse features of stancetaking language as it appears in this corpus.

3.3.1 The ICLE corpus

Collected by Granger (2003) and originally intended for use by second language acquisition researchers, ICLE is a collection of largely argumentative essays written by non-native speakers of English, each essay responding to one of 14 separate essay prompts. We chose this corpus for the current study because of its size, diversity of topics, and availability. Other student essay resources, such as Educational Testing Service's collection of standardized essay responses, the Michigan Corpus of Upper-Level Student Papers (Ädel and Römer, 2012), The International

ESSAY PROMPT	PROMPT ABBREV.
All armies should consist entirely of professional soldiers: there is no value in a system of military service.	*Armies*
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	*Degrees*
Feminists have done more harm to the cause of women than good.	*Feminism*
Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.	*Marx*
In the words of the old song, “Money is the root of all evil.”	*Money*
The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.	*Prisons*
Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?	*Science*

Table 3.1: List of ICLE prompts used in the current study. For ease of reference, a mnemonic for each prompt is given in the second column.

Corpus Network of Asian Learners of English (Ishikawa, 2011), and the Upsala Student Essays corpus (Axelsson, 2000), are either too restrictive in their permitted research use, too small for the purposes of this study, or contain too few essay topics.

After reviewing the 14 essay prompts provided to students, the seven topics given in Table 3.1 were chosen. In the second column, we have included the mnemonics we will use to reference each prompt throughout this study.

3.4 Stancetaking language versus attitudinal language

Recent NLP work in Sentiment Analysis makes a distinction between the tasks of opinion identification/classification, which deals with the automated detection of opinion-bearing language and the classification of that language as positive or negative, as described in section 2.2, and stance identification/classification which identifies stancetaking language and classifies that language as taking a positive (=*for*) or negative (=*against*) stance. This distinction can be partly maintained based on evidence from annotation studies (Wilson, 2008) showing clear speaker intuitions regarding a difference between language that expresses "positive (negative) sentiment" and language that expresses "positive (negative) arguing" (Wilson, 2008, pg. 126). Additionally, quantitative analyses (Hunston, 2007) comparing the vocabulary used in text containing the highly emotional language of opinion, such as the tabloid news reportage of the *Sun* newspaper, and text containing stancetaking language, such as the science periodical *New Scientist*, show distinct differences in the kinds of adjectives employed in particular lexico-syntactic patterns across these two domains. The sentences given in (8) both contain instances of the pattern *it-v-link-ADJ-that* which are are typical of writers' markedly different adjective choices in these two text varieties.

(8) a. It is **scandalous** that the rich can buy the drugs privately, but tough luck if you are poor. (*Sun*) —from Hunston (2007, pg. 34)

b. [...] it was **possible** that strontium and calcium in fossils might have reacted chemically with the rock in which the fossils were buried. (*New Scientist*) —from Hunston (2007, pg. 34)

Use of *scandalous* in (8a) allows the writer to comment on the morality of *the rich can buy drugs privately, but tough luck if you are poor* while use of *possible* in (8b) hedges the likelihood of *strontium and calcium in fossils might have reacted chemically with the rock in which the fossils were buried*. Negative adjectives of social sanction, such as *scandalous, avaricious*, and *devious* are categorized as *judgment* lexemes in Martin and White's taxonomy of evaluative language which in turn are grouped under the higher-level category of attitudinal language. Martin and White (2005) place *possible* under the rubric of *entertain* vocabulary, itself a sub-category of the language of *engagement*. The language of *engagement* allows speakers to comment on the truth or likelihood of an accompanying proposition—this is roughly equivalent to what we have been calling stancetaking language.

In line with Martin and White's lexical taxonomy of evaluative language, we call the emotion-laden language of opinion *attitudinal language*. This is the kind of evaluative language that has served as the test bed for the majority of NLP research dealing with the classification of evaluative language. Attitudinal language is concerned with registering a writer's feelings towards an entity or behavior and is typically an adjectival phenomenon, as the boldfaced terms in (4c), repeated below as (9a) and (9b), taken from a newspaper opinion piece, show.

(9) a. "Snake Eyes" is the most **aggravating** kind of movie: the kind that shows so much **potential** then becomes **unbelievably disappointing**. —From Pang et al. (2002)

b. Hollande certainly has his **flaws**. A deliberately **unassuming** style of leadership has allowed factions to flourish and, at times, confusion to reign. In deliberately avoiding the **hyperactive** approach to the presidency that characterized the Sarkozy years, he has often appeared too **laid back, humble** and **sanguine** for his own good. —Julian Coman, the Guardian, 4/27/13

The boldfaced terms in (9a,b) display all of the lexical hallmarks of attitudinal language. In such cases, the writer's choice of lexis is usually motivated by the entity being evaluated. In (9a), the entity is the movie *Snake Eyes*, while in (9b) it is the behavior of François Hollande. Martin and White divide attitudinal lexis into discrete categories depending on the type of entity the attitudinal word is meant to describe. Cultural artifacts such as books and movies call for a vocabulary of *appreciation*, which deals with aesthetic evaluation (*aggravating*, *disappointing*), while the evaluation of behavior calls for a vocabulary of *judgment* (*unassuming, hyperactive, laid back, humble, sanguine*). A given example of opinionated text will contain *appreciation* and *judgment* lexis in varying proportion. Readers' intuitions regarding different genres of opinionated text—the difference between a movie review and an editorial, for example—can often be ascribed to the dominance of one category of lexis relative to another. In movie reviews, for example, we expect to find *appreciation* dominant while *judgment* will likely dominate in an editorial.

Judgment lexis is used freely in the ICLE corpus, particularly when the argument involves controversial socio-political issues such as the perceived dominance of technology (10a) or the legacy of feminism (10b):

(10) a. There is a **civilized cruelty** in the world of science, in the **cold, calculated** interest of the probing eyes and shining scalpels, the childhood dream lying **helpless** upon a bed of **sterile** white. —*Science* prompt

b. Boys are rather **rational**, girls more **emotional**, boys are usually more **ambitious** or **wilder** than girls who are very often **placid** and **modest** as far as their career is concerned. —*Feminism* prompt

Judgment lexis can take a number of forms, depending on the type of behavior being evaluated. In (10a), the writer buttresses her argument that science and technology have displaced imagination by describing the behavior of scientists using a vocabulary of negative social sanction: Scientists display a *civilized cruelty*, and *cold, calculated interest*. In (10b), on the other hand, words such as *rational, ambitious, placid, modest, emotional* and *wilder* serve to express the positive or negative "social esteem" (Martin and White, 2005, pg. 52) in which an individual or group should be held.

In contrast to (9) and (10), markers of stancetaking tend to be more grammatically diverse than those of attitudinal language as the ICLE corpus sentences in (11) illustrate.

(11) a. Women **should** be treated equally as men and these does not mean that women should be given tasks that are beyond their control. —*Feminism* prompt

b. I **believe** that only imagination and dreaming are able to make the world go round. —*Science* prompt

c. One thing I'm sure of: crime will **certainly not** decrease by building more prisons. —*Prisons* prompt

As we will discuss in section 3.6, stance is most commonly expressed across and within different grammatical classes such as verbs (both lexical and modal) and adverbs. In (11a), the necessity (or, deontic) modal *should* is used to advocate the moral necessity of the proposition *women BE treated equally*. In (11b), the writer's commitment to the proposition *only imagination and dreaming are able to make the world go round* is hedged using *believe*, one of a group of epistemic judgment verbs (verbs showing speaker commitment to the truth of the accompanying proposition) that also includes *suggest, indicate, speculate*, and *assume*. Finally, in example (11c), the modal adverbial *certainly not* allows the writer to comment on the degree of certainty associated with the proposition being evaluated (Biber et al., 1999, pg. 972). In this case, the likelihood of the proposition *crime decrease by building more prisons* is evaluated as highly doubtful.

Aside from lexical differences in the vocabulary used to express attitude and stance in (10) and (11), these two groups of sentences also exemplify an important semantic distinction between the sorts of objects that writers evaluate using attitudinal language versus stancetaking language. The targets of the attitudinal adjectives in (10) are entities, such as *cruelty* and *interest* in (10a) and *boys* and *girls* in (10b) while the targets of stancetaking language in (11) are full propositions. In the next section, we discuss this important distinction further.

3.5 Semantic features

An influential view in the literature on evaluative language (summarized in Martin and White, 2005, pg. 38) proposes that the key distinction between attitudinal and stancetaking language turns on the semantic class of the target of the attitude or stance with attitudes taking entities and stance taking propositions as targets. This line of research begins with the syntactic observation (Hunston and Thompson, 2000, pg. 3) that the quintessential markers of stance in English, such as verbs of epistemic judgment (*think, believe*), necessity (or, deontic) modals (*ought, should*), and modal adverbs (*possibly, certainly*), all fall within grammatical classes that typically select for full clauses rather than noun phrases. Since the traditional semantic role associated with the clause is the proposition (Tarski, 1944, pg. 342) while the role associated with the noun phrase is the entity (Chierchia and MacConnell-Ginet, 2000, pg. 96), it is reasonable to propose an opinion=entity target/stance=proposition distinction to capture these syntactic facts semantically. The distinction is also implicit in Halliday and Mathiessen's (1994) comment regarding the status of propositions as targets of verbs of epistemic judgment:

> When language is used to exchange information, the clause takes on the form of a **proposition**. It becomes something that can be argued about—something that can be affirmed or denied, and also doubted, contradicted, insisted on, accepted with reservation, qualified, tempered, regretted and so on. (Halliday and Matthiessen, 2004, pg. 110)

In other words, for Halliday and Mathiessen, only propositions can be affirmed, denied, doubted, and so on for other stancetaking verbs taking clausal complements. A similar claim can be made for any of the other grammatical classes of stancetaking markers we have mentioned. Only propositions can be a possibility or a certainty, or endorsed to be the case using *should* or *ought*. By contrast, entities such as books or films cannot be affirmed, denied, doubted, etc.

The attitude=entity/stance=proposition distinction has important consequences for the kinds of corpora resources used in NLP tasks involving attitudinal language versus those used in classification tasks involving stancetaking language. In general, opinion (attitude) classification tasks deal with review corpora which take cultural artifacts (i.e., entities) such as movies, books, and gadgets as their targets (Tang et al., 2009) while stance classification tasks have concentrated on debate corpora which take full propositions as their targets. The online debates used as corpora in Somasundaran and Wiebe (2010), Walker et al. (2012a), Hasan and Ng (2013a), and Hasan and Ng (2013b) are headed by topic posts such as *God exists*, *Abortion should be banned*, and *Creationism is false*. These topic posts serve, in their original or reworded form, as the target propositions of stance markers in debate posts arguing *for* or *against* the statement in the topic post, as shown in (12-14).

(12) a. God exists.
b. God absolutely does not exist and that is why there are so many wrongs in the world.

(13) a. Should abortion be legal?
b. Abortions have to be legal or all hell will break loose.

(14) a. Should marriage for same-sex couples be legal?
b. I cannot agree with gay marriage because I believe in right and wrong and that homosexuality is wrong.

In (12b), the writer's negative stance involves simply reproducing the proposition in the topic post (12a) and modifying that proposition with *absolutely does not*. Examples (13) and (14) are more structurally complex, in part because the topic posts themselves are not declaratives and so the underlying propositions must be reconstructed by the writer using declarative inversion. In (13b), the reconstructed proposition in (13a), *Abortion should be legal*, serves as the target of *have to*. In (14b), a version of the reconstructed proposition *Marriage for same-sex couples should be legal* in (14a) is argued against using *cannot agree*.

The sentence pattern observed in (12-14) is rare in the ICLE corpus, though (15a) is one such example. In general, students tend to avoid the tactic of expressing their stance toward the proposition given in the administered essay prompt by simply reproducing, in whole or part, that proposition along with some sort of stance marker indicating whether they are *for* or *against* that proposition. As we will show, students express sentence-level stance using a far greater variety of lexico-syntactic resources than occurs in online ideological debates. When the pattern seen in (12-14) does occur in the ICLE corpus, the proposition taken from the topic prompt tends to be referenced exophorically, as in (15b), where *this statement* references the prompt statement.

(15) a. The claim that there is no longer a place for dreaming and imagination because our modern world is dominated by science technology and industrialization is in my opinion downright stupid.
b. In my opinion this statement is rather doubtful although there are certain reasons for some people to think so.

3.6 Lexico-syntactic features

Much of the research dealing with the lexico-syntactic features of stancetaking language is focused on those features that can be collectively desribed as "evidential." Evidential markers allow speakers to express certainty or doubt concerning the truth of a proposition. These markers include the classes of lexical verbs (*conclude, demonstrate, indicate*), modal adverbs (*assuredly,* *indeed, allegedly, supposedly*), hedges (*perhaps, maybe*), boosters (*certainly, clearly*), predictive modals (*will, would*), possibility modals (*might, could*), and necessity (or, deontic) modals (*ought to, should*). For some researchers (Chafe and Nichols, 1986; Biber and Finegan, 1989), the relationship between stancetaking language and evidentiality is so close that the two terms can be used interchangeably. Halliday and Matthiessen (2004), for example, claim that evidentiality "enacts the speaker's opinion—an enactment of his or her degree of commitment to the proposition [. . .]" (Halliday and Matthiessen, 2004, pg. 605). In this short description, the reader will recognize many of the features of stancetaking language that we considered in sections 3.4 and 3.5. A distinction is made between stancetaking and attitudinal language (section 3.4)—the former serves to enact the latter—and the target of the stance is assumed to be a proposition (section 3.5).

Evidentiality is ubiquitous in the ICLE corpus. In (16-22), we can observe several representative types of evidentiality found in the ICLE corpus along with example sentences. Selected evidentials are underlined and their target propositions are boldfaced.

(16) **Verbs of epistemic judgment**

a. This indicates that **our prisons are higher institutions for criminals.**
—*Prisons* prompt

b. Personally, I suggest that **our government should make some laws and regulations to force every young man who is not invalid to hold 'Military Service in certain age.** —*Armies* prompt

(17) **Modal adverbs**

a. The prison system is used to punish criminals supposedly **to stop them from committing crime after serving sentence but this hardens them instead.** —*Prisons* prompt

b. **You can't** possibly **decide when someone should get capital punishment.** —*Prisons* prompt

(18) **Hedges**

a. Perhaps **they are buried deeper somewhere in the depths of the subconscious.**
—*Marx* prompt

b. Maybe **the people of today have lost some of that companionship.**
—*Marx* prompt

(19) **Boosters**

a. **Television is** certainly **people's Bible as far as information processing is concerned.** —*Prisons* prompt

b. **Science technology and industrialization** clearly **dominates the world we live in.** —*Science* prompt

(20) **Predictive modals**

a. When you are going to present yourself for a job interview you know you are not going to be chosen by your curriculum but you will be employed or not depending on what you tell in the interview and sometimes **you** will **have to answer ridiculous questions.** —*Degrees* prompt

b. **Those things** would **be very interesting to learn in university.** —*Degrees* prompt

(21) **Possibility modals**

a. Another answer might be that **the Norwegian politicians have not faced the fact that Cold-War is over.** —*Armies* prompt

b. I also agree that **rehabilitation** could **be a nice alternative.** —*Prisons* prompt

(22) **Deontic modals**

a. In short I agree with the Ancients: **the "happy mean"** should **be our real ideal the, path to be followed.** —*Money* prompt

b. **Something** ought **to be done to improve the prison system** and the way to rehabilitate criminals should be found. —*Prisons* prompt

The evidential markers used in (16-22) calibrate writer certainty concerning the likelihood of the accompanying proposition. Importantly, using an evidential to downgrade or upgrade the likelihood of the accompanying proposition does not have the effect of reversing the stance polarity (*for* or *against*) of that proposition. Use of *maybe* in (18b) to downgrade the likelihood of the positively stanced proposition *people of today have lost some of that companionship* does not indicate to the reader that the writer is arguing against that proposition—the truth of the proposition has simply been hedged. Similarly, use of *possibly* in (17b) to downgrade the likelihood of the negatively stanced *you can't decide when someone should get capital punishment* does not turn this into a positively stanced proposition. We will encounter many of these forms again in chapter 5 when we describe our method of feature engineering for the stance classification experiments reported in chapter 6.

3.6.1 The syntactic relationship between stancetaking language and propositions

We have described the influential view that stancetaking language is used to express a speaker or writer's degree of commitment toward a proposition. As mentioned in section 3.5, the syntactic realization of the proposition is the clause. These two observations have led researchers in stancetaking language to focus on the clause-selecting lexis featured in (16-22) as the central vocabulary of stancetaking. A key aspect of the feature set described in section 5.2 involves the extraction of stance words falling into the classes given in (16-22), along with the propositions targeted by those stance words. It is important, therefore, that we determine the general relationship between stancetaking lexis and the clausal proposition targeted by that lexis. Since we are primarily interested in describing the clausal behavior of stancetaking language, we make use of a shallow syntactic representation in our syntactic descriptions. In the examples provided below, all sentences have been syntactically parsed into Penn Treebank format using the Stanford parser (Klein and Manning, 2003). All stancetaking words have been identified in these examples using the stance lexicon constructed in Somasundaran (2010) and Somasundaran and Wiebe (2010) (the details of this lexicon are described in section 5.2.5).

Our first example involves stancetaking lexis targeting a proposition in an adjacent clause. In the simplest scenario, this takes the form of an epistemic judgment verb (*indicate, think*) taking a *that*-clause object, which in turn contains the proposition being evaluated. We see this in example (16a), reproduced below in parsed format as (16a)′.

(16a)′ [$_{ROOT}$ [$_{S}$ [$_{NP}$ This] [$_{VP}$ [$_{VBP}$ **indicates**] [$_{\bar{S}}$ [$_{IN}$ that] [$_{S}$ our prisons are higher institutions for criminals.]]]]]

In *that*-clause object constructions, such as (16a)′, the S-clause encompassing the target proposition can be found by simply tracing a path downward to the S-node that next occurs after the appearance of the stance word.

When a stance is taken on a proposition in the immediate clause, this often involves adverbials such as the hedge *perhaps* or the booster *certainly* which comment on the likelihood of that proposition. Scoping comment adverbials can be left *in situ* or incorporated into the proposition, as in (18a) and (19a), respectively, given below as (18a)$'$ and (19a)$'$.

(18a)$'$ $[_{ROOT}$ $[_{S}$ $[_{ADVP}$ **Perhaps**] $[_{NP}$ computers] $[_{VP}$ $[_{VB}$ are] today's most important invention] $[_{S} \ldots]$]]

(19a)$'$ $[_{ROOT}$ $[_{S}$ $[_{NP}$ Television] $[_{VP}$ is $[_{ADVP}$ **certainly**] people's bible as far as information processing is concerned.]]]

Other cases in which stancetaking lexis and the proposition targeted by that lexis are integrated into the same clause include those sentences containing the various classes of epistemic modals given in (20-22). In such cases, modals "create an irrealis scope over the proposition in which they are lodged." (Givón, 1993, pg. 171). The relevant clauses of these examples have been reproduced below.

(20b)$'$ $[_{ROOT}$ $[_{S}$ $[_{NP}$ Those things] $[_{VP}$ $[_{MD}$ **would**] $[_{VP}$ be very interesting to learn in university]]]]

(22a)$'$ $[_{ROOT}$ $[_{S} \ldots$ $[_{S} \ldots$ $[_{S}$ $[_{NP}$ the happy mean] $[_{VP}$ $[_{MD}$ **should**] $[_{VP}$ be our real ideal]]]]]]

(22b)$'$ $[_{ROOT}$ $[_{S}$ $[_{NP}$ Something] $[_{VP}$ $[_{MD}$ **ought to**] $[_{VP}$ be done to improve the prison system] $[_{CC}$ and] $[_{S}$ $[_{NP}$ the way to rehabilitate criminals] $[_{VP}$ **should** be found]]]]]

The S-clause encompassing the target propositions in (20b)$'$, (22a)$'$, and (22b)$'$ is the immediate S-clause of the identified stance word. In (20b)$'$ and (22b)$'$, this S-node is the sentence itself while in the (22a)$'$ it is lowermost of three embedded S-nodes: $[_{S}$ *the happy mean **should** be our real ideal].*

For each of the modals in (20b)$'$, (22a)$'$, and (22b)$'$, we can more clearly indicate their role in taking irrealis scope over the proposition in the immediate clause by representing each sentence as a one-place predicate with the modal taking the proposition as a complement, as in (23). (This representation is adopted from the classic analysis of epistemic modality of Ross, 1969).

(23) a. should (the happy mean be our real ideal)

b. would (those things be interesting to learn in university)

c. ought to (something be done to improve our prison system)

The examples we have considered so far suggest a fairly regular pattern in which the proposition evaluated by a stance marker is located in the immediate or embedded clause. But does this pattern also apply to those cases of doubly embedded instances of stancetaking, as in (16b) and (21b) reproduced below as (16b)′ and (21b)′? In (16b)′, the matrix clause includes the stance marker *suggest* which takes the *that*-clause object, *our goverment should make some laws and regulations to force every young man who is not invalid to hold "Military Service" in certain age*. The latter, in turn, includes the possibility modal *should* which scopes over the proposition in which it is lodged. We see a similar embedded pattern in (21b)′with *agree* taking a *that*-clause object, *rehabilitation could be a nice alternative*. The latter contains the embedded possibility modal *could* which scopes over *rehabilitation be a nice alternative*.

(16b)′ [$_{ROOT}$ [$_{S}$ [$_{AdvP}$ Personally] [$_{NP}I$] [$_{VP}$ [$_{VBP}$ **suggest**] [$_{\bar{S}}$ [$_{IN}$ that] [$_{S}$ [$_{NP}$ our government] [$_{VP}$ [$_{MD}$ **should** make some laws and regulations to force every young man who is not invalid to hold "Military Service" in certain age.]]]]]]]

(21b)′ [$_{ROOT}$ [$_{S}$ [$_{NP}$ I] [$_{AdvP}$ also] [$_{VP}$ [$_{VBP}$ **agree**] [$_{\bar{S}}$ [$_{IN}$ that] [$_{S}$ [$_{NP}$ rehabilitation] [$_{VP}$ [$_{MD}$ **could**] [$_{VP}$ be a nice alternative]]]]]]]

Maintaining the pattern observed above, with a stance marker evaluating a proposition in the immediate or embedded clause, both *suggest* and *should* would take the proposition *our government make some laws and regulations to force every young man who is not invalid to hold "Military Service" in certain age* as their target. The same iterated structure would hold for (21b)′ with both *agree* and *could* taking *rehabilitation be a nice alternative* as a target. These structures, with each stance word targeting the same proposition, can be represented as the one-place predicates in (24).

(24) a. suggest (our goverment make some laws and regulations to force every young man who is not invalid to hold "Military Service" in certain age.)

b. should (our government make some laws and regulations to force every young man who is not invalid to hold "Military Service" in certain age.)

c. agree (rehabilitation be a nice alternative)

d. could (rehabilitation be a nice alternative)

The representation given in (24) makes the implicit claim that the embedded modal operators in (16b)′ and (21b)′ are part of the same stancetaking act as the attitude verbs contained in the matrix clause. This in turn means that the embedded modals are anchored to the same subject as the matrix verb. We find support for this claim in research dealing with epistemic modals embedded in sentences containing attitude or belief verbs. In the earliest such research, Antinucci and Parisi (1971) propose that both *says* and *must* in sentence (25) are anchored to the attitude holder *John* since the default reading of epistemic modals is always egocentric—i.e., tied to the speaker.

(25) John says that Harry must have gone.

The proposal of Antinucci and Parisi has been revived in recent debates (Hacquard, 2006; Anand and Hacquard, 2008) concerning epistemic modals embedded under belief reports, as in the example of Hacquard (2006, pg. 139, ex.213a) and Anand and Hacquard (2008, pg. 4, ex.6b), adapted below as (26).

(26) Darcy believes that it might be raining.

a. *Rain was possible for Darcy at her past (believing) time*

b. *# Darcy believed that rain* {*was/is possible for me, possible for him*}

Under the analysis of Hacquard (2006) and Anand and Hacquard (2008), when a modal is used as the complement of an attitude, as it is in (26), it triggers a default reading, given in (26a), in which that modal is tied to the attitude event and, by extension, to the holder of that attitude. This explains the odd-sounding reading given in (26b) in which the modal is tied to another attitude holder. Additionally, as shown in Hacquard (2006, pg. 140, ex.214), given below as (27), the default egocentric reading of epistemic modality explains the seeming redundancy of (27a), which is truth conditionally equivalent to (27b).

(27) a. ?I believe it might be raining.

b. It might be raining.

Our iterated representation of embedded epistemic modals given in (24), with the matrix verb and modal both scoping over the proposition, captures these insights—both stance markers are tied to a single proposition (the “attitude event”) which is in turn tied to the holder of that proposition. This is a simple representation that can easily be implemented as a classification feature.

Summing up, the propositions targeted by stance words taking *that*-clause objects are located in the embedded S-clause of that word. Otherwise, all propositions are located in the immediate S-clause of a given stance word. In those cases involving doubly embedded stance markers, the stance markers and the proposition have a many-to-one relationship with each stance marker evaluating the same proposition. This is the pattern that we assume when constructing our feature set in section 5.2.

3.6.2 Pollyanna effects in stancetaking language

The claim of Boucher and Osgood’s (1969) Pollyanna Hypothesis is that there is a universal human tendency to use positive rather than negative language, a tendency reflected in the highly skewed distribution of positive versus negative words in the world’s languages. Recently, there has been a revival of interest in the Pollyanna Hypothesis in the context of opinion mining research.

The distributions of positive versus negative opinion-bearing words in the lexicons collected for use in several opinion mining systems appear to validate this hypothesis. Mohammad et al. (2009), for example, report that the percentage of positive and negative opinion polarity items in their lexicon is 65% and 35%, respectively, and the MPQA subjectivity lexicon (Wilson and Wiebe, 2005) has a nearly identical distribution of 64% positive and 36% negative words.

Do we also find Pollyanna effects in stancetaking language? To answer this question we can examine available collections of stancetaking lexis divided into *for* and *against* polarities. For example, in the stance lexicon constructed by Somasundaran and Wiebe (2010) and Somasundaran and Wiebe (2010) we find that fully 82% (3094) of terms are marked *for* while only 18% (668) are marked *against*. Looking at a second collection, Martin and White's (2005) taxonomy of evaluative lexis, we find a similar bias towards positively stanced language. In their analysis of lexical markers of *engagement*, a category of evaluative language that includes the "linguistic resources by which speakers/writers adopt a stance towards the value positions being referenced by the text" (Martin and White, 2005, pg. 92), we find a marked emphasis on positively stanced words. Of the nine sub-categories of *engagement* described, only one involves lexical markers that allow speakers/writers to argue against a proposition. This is the category of *deny* words, such as *no*, *didn't*, and *never*. The focus on positively stanced lexis in these two resources suggests that Pollyanna effects occur in stancetaking language in much the same way that they occur in opinion-bearing language.

In the ICLE corpus, Pollyanna effects are evident in the tendency among essay writers to express an *against* stance by simply negating a positively stanced word, as in (28).

(28) a. Some even argue that musicians **should not** need to go to school to develop a high level of proficiency on their instruments. —*Degrees* prompt

b. Science technology and industrialisation **must not** dominate our world. —*Science* prompt

c. But Marx would **probably not** have replaced religion by television because the nineteenth century is much too different from the twentieth and things have changed rapidly. —*Marx* prompt

d. One thing I'm sure of, crime will **certainly not** decrease by building more prisons. —*Prisons* prompt

We will revisit both the Pollyanna Hypothesis and the use of negated *for* lexis to negatively evaluate a proposition in section 5.2.

3.7 Discourse features

The features we have considered so far have mostly dealt with the linguistic characteristics of stancetaking language as they occur inside the essay response. They are concerned with what writers *say* when they use stancetaking language. However, it can be argued that paralinguistic features play an equally important role when writers take a stance. When considering this role, we are concerned with what writers *do* when they take a stance on the proposition contained in the prompt statement. To whom or what is the stancetaking language directed? Is the act of stancetaking simply an asymmetric relationship between writer and proposition or is there another writer or proposition involved? We can answer these questions by identifying the discourse features of stancetaking language as they occur in the ICLE corpus.

Discourse features in the ICLE corpus can be divided into two distinct categories: *endophoric* (inside the text) discourse features, which link different sections of the text, and *exophoric* (outside the text) features which link sections of the text to the prompt statement. The most prominent endophoric features in the ICLE corpus are based on organizational conventions used to guide the reader through the text such as *I will show*, *in this section*, and *in addition*. These markers are so ubiquitous in stancetaking discourse that researchers have accorded them a central role when trying to understand the intersentential resources used by writers of stancetaking genres such as academic research articles and argumentative essays.

The term *metadiscourse* was introduced in Harris (1958) to describe segments of scientific articles that "talk about the main material" and guide the reader's interpretation of that material. It is this narrow conception of the term that influenced Williams' (2007) discussion in his well-known style guide. Williams describes metadiscourse as those words and/or phrases that allow writers to comment on their own thinking (*I will show/argue/claim*), directly address readers (*consider now, as you will recall*), or provide a logical connection (*first, therefore*). Subsequent research on metadiscourse in academic writing (Kopple, 1985; Crismore and Farnsworth, 1989; Hyland and Tse, 2004) broadened Harris' analysis. In this research, metadiscourse includes signposts of rhetorical organization that guide the reader through the text and markers of language that allow writers to evaluate propositional content via attitudinal language (*It is alarming to note, surprisingly*) and epistemic language such as hedges (*might, perhaps*). Though most commonly applied to the analysis of academic prose, the metadiscourse framework has been used to analyze other genres, including Darwin's *The Origin of Species* (Crismore and Farnsworth, 1989), school textbooks (Hyland, 2004), and company annual reports (Hyland, 1998).

Hyland's (2005) model emphasizes the role that stancetaking language plays in expressing reader-writer interactions. When authors insert themselves into a text, they do so in order to directly engage the reader using the organizational and evaluative resources of metadiscursive language. Following Thompson and Thetela (1995), Hyland describes these as the *interactive* modes of metadiscourse (another set of markers, termed *interactional*, are nearly identical to those we described in section 3.6 and so are ignored here). Interactive language includes all those linguistic expressions that help to guide readers through the text. This includes, among others, transitions (*in addition, thus*), frame markers (*finally, to conclude*), and evidentials (*according to X, as X concludes*). This model of metadiscourse is compatible with our assumption that there is a distinction between the propositional content of a given stancetaking text and linguistic markers that signal the author's stance toward that content or serve to organize it in some way. Hyland's conception of metadiscourse also makes a distinction, discussed in section 3.4, between attitudinal language, which Hyland calls “attitude markers,” and stancetaking language (for Hyland, these are modals and hedges).

In the ICLE corpus, metadiscourse is often used to comment on an essay’s overall structure, as in (29).

(29) a. **I will first introduce** my as objective as possible vision of the university and of its most important aspects and then I will argue on the sentences which form the title of this essay. —*Degrees* prompt

b. **To conclude** we should acknowledge that although television is the new opium of the masses it remains for many people especially old people a kind of companion. —*Marx* prompt

In (29), illocution markers (Hyland, 2005, pg. 32) such as *I will first introduce* and *to conclude* allow the writer to explicitly guide the reader’s interpretation of the writer’s intent at key points in the essay. Given the relatively short length of the standardized essay response, however, it is more common for metadiscourse in this context to occur at the rhetorical boundaries of supporting arguments, as in (30).

(30) a. The church will be justifying the numerous taxes obedience to the rule and so forth by introducing the notion of divine right and of respect of leadership. **Moreover** to forget or accept the mediocrity of their situation people took refuge in praying and hoping for a better place in paradise. —*Marx* prompt

b. The main intention of imprisonment is to protect the society from serious crime in general and dangerous criminals in particular. **In addition** it is also meant to prevent people from committing criminal actions. —*Prisons* prompt

Conjunctive adverbs such as *moreover* and *in addition*, both of which are categorized by Hyland as logical connectives, serve to elaborate upon and to reinforce the argument offered in the previous sentence.

The unique writing context of the student argumentative essay gives rise to discourse effects that also appear in other argumentative exchange settings such as political debates and online ideological debates. Exophoric discourse effects in the ICLE corpus are a consequence of the relationship between the essay prompt and the essay response. That there exists a relationship between prompt and response is obvious from the lexical overlap evident in prompt/response pairs such as (31), where (31a) is the text of the *University degrees* essay prompt and (31b) is the first sentence of a response to that prompt.

(31) a. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.
b. Nowadays there have been many debates on whether most university degrees are theoretical and don't prepare students for real world or not.

Explicit exophoric reference to the prompt statement is also common, as in (32).

(32) I find this statement very true indeed.

For Du Bois (2007), exophoric discourse effects—in particular, those that occur as part of a conversational exchange—can be represented as vertices and directed edges on a graph or "stance triangle." The stance triangle structure captures both the subjective relationship between a speaker and the proposition that the speaker is evaluating, and also the intersubjective relationship between the speaker and another speaker with whom the first speaker aligns or disaligns herself when she takes a stance on a proposition. The vertices of the triangle are the two subjects and the proposition being evaluated. In the act of taking a stance on the proposition, the first subject aligns or disaligns herself with the second subject: "Alignment can be defined [...] as the act of calibrating the relationship between two stances, and by implication between two stancetakers."(Du Bois, 2007, pg. 144)

We have adapted those aspects of Du Bois's model that are most relevant to the kind of stancetaking that occurs in the context of the standardized essay response (ignoring Du Bois's category of *positioning*, which is redundant here). In the diagram given in Figure 3.1, Subject 1 can be

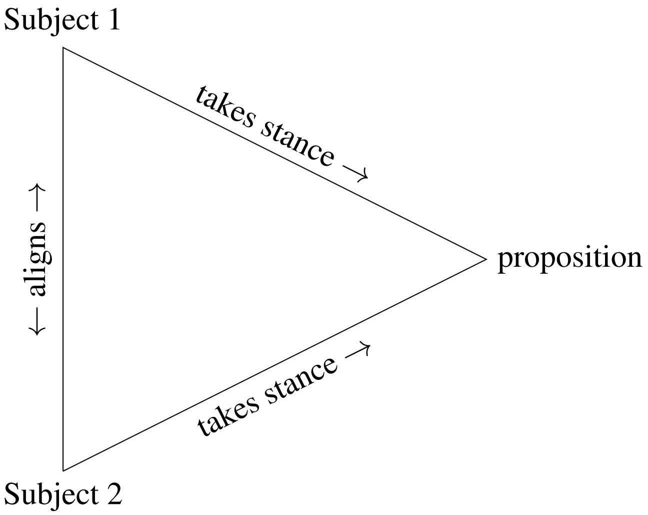


Figure 3.1: Du Bois's stance triangle. Subject 1 and Subject 2 are exchange participants; the proposition serves as the target of the exchange. By taking a stance on the proposition, Subject 2 aligns (disaligns) herself with Subject 1.

viewed as the author of the essay prompt, Subject 2 can be viewed as the author of the essay, and the proposition can be viewed as the statement taken from the essay prompt.

The stance triangle model is implicit in both the Metadiscourse and Appraisal Theory (Martin and White, 2005) accounts of evaluative language. For Hyland (2005) and Hyland and Tse (2004), the evaluation of an object involves the interaction of two subjects, the writer and an imagined reader. Martin and White follow Halliday and Matthiessen (2004) in locating evaluative language within the "interpersonal" dimension of language use.

Given Du Bois's focus on stancetaking as an intersubjective act, it is not surprising that he relies on data drawn from a corpus of conversational exchanges (Du Bois, 2000) to illustrate the key features of his model of interactional stancetaking. When applying Du Bois's model to the ICLE text, we assume that the prompt writer and the essay writer are exchange participants. In (33) and (34), both prompt text and essay response text are presented as an exchange with stance (*for* or *against*) and alignment values (aligned or disaligned) provided below each sentence.

(33) a. **Prompt author (Subject 1)**: The prison system is outdated.

Stance: *for*
Aligned or disaligned? n/a

b. **Essay author (Subject 2):** Initially I would say that the prison system is old-fashioned, especially when it comes to the rehabilitation of criminals.
Stance: *for*
Aligned or disaligned? Aligned

(34) a. **Prompt author (Subject 1):** Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination.
Stance: *against*
Aligned or disaligned? n/a

b. **Essay author (Subject 2):** No matter how modern the world becomes there will always be room for dreaming and for imagination.
Stance: *for*
Aligned or disaligned? Disaligned

In (33), the prompt writer and essay writer are aligned in their *for* stance toward the proposition, *The prison system is outdated*. A *for* stance polarity is maintained using the *for*-stancetaking expression *would say* along with a complement clause containing a synonym of the sentiment-bearing word contained in the proposition matched for sentiment polarity: *old-fashioned* is a synonym of the negative sentiment-bearing *outdated*. By contrast, in (34), the prompt writer takes a stance *against* the proposition *in our modern world, dominated by science, technology, and industrialization, there is a place for dreaming and imagination*. Disaligning herself with the prompt author involves taking a *for* stance toward the proposition using the predictive modal *will* and a paraphrased version of the original proposition, *there BE always be room for dreaming and imagination*. What is clear from the exchanges in (33) and (34) is that stancetaking language in prompt-essay response exchanges is often realized by selectively reproducing segments of the target proposition and thereafter modifying those segments with appropriate expressions of stancetaking language.

3.8 Summary

In this chapter, we presented a set of descriptive generalizations that capture the core characteristics of stancetaking language at the semantic, lexico-syntactic and discourse levels. These linguistic generalizations will motivate our creation of classification features in chapter 5.

In sections 3.4 and 3.5, we distinguished the targets of stancetaking language from those of attitudinal language by noting that stance targets tend to be clausal while attitude targets tend to be nominal. The fact that the semantic roles generally associated with these syntactic categories are the proposition and entity, respectively, allows us to make the semantic generalization that expressions of stance take propositions as targets while those of attitude take entities as targets. This semantic distinction has lexico-syntactic consequences for the sorts of expressions we expect to find modifying targets of stance versus targets of attitude. In keeping with its role of describing the positive or negative characteristics of nominal material (an entity), the prototypical grammatical class of attitudinal language is the adjective. By contrast, as discussed in section 3.6, the expressions most often associated with stancetaking are evidential—verbs of epistemic judgment, modals, hedges, and boosters—and generally select for full clauses (propositions).

In section 3.6.1, we found that stancetaking words target propositions in the immediate or emebedded clause, a pattern that also holds for instances of doubly embedded stance markers. To support the latter claim, we drew on the recent work of Hacquard (2006) and Anand and Hacquard (2008), which deals with sentences containing epistemic modals embedded in belief reports. The default reading of such modals is anchored to the subject of the attitude verb in the matrix clause. We can capture this insight by representing both the attitude verb in the matrix clause and the epistemic modal in the embedded clause as targeting the proposition in the embedded clause.

In section 3.7, we turned our attention to the discourse-level features of stancetaking language. We first discussed the organizational, or metadiscourse cues used by essay writers to guide their reader through their arguments. Given the length constraints imposed, ICLE essay writers make limited use of illocution markers such as *I will argue* and *to conclude*, but make frequent use of logical transition markers such as *moreover* and *in addition*.

We explained the lexical overlap between prompt and response evident throughout the ICLE corpus by drawing on the stance triangle research of Du Bois (2007). Under the stance triangle model, essay authors align or disalign themselves with the author of the prompt statement by selectively reproducing segments of the proposition contained in that statement and then arguing either *for* or *against* that proposition using appropriate expressions of stance.

Chapter 4

Essay-level stance annotation

In this chapter, we will describe the process of creating a corpus of ICLE essays annotated for document level stance. The corpus creation work presented here adopts standard methodologies of document-level sentiment annotation (O'Hare et al., 2009; Macdonald and Ounis, 2006), but we take a crowdsourced approach to this task. As Mellebeek et al. (2010) report, the quality of crowdsourced document-level annotation approaches that of expert annotation and involves far less time and expense. The annotations described in this chapter will be used in the supervised document-level stance classification experiments reported in chapter 6.

4.1 Background

As described in section 2.1, the paradigmatic example of sentiment-bearing language is the opinionated, emotion-bearing language of reviews. Thus, the earliest modern studies of sentiment analysis, that of Das and Chen (2001) and Pang et al. (2002), use corpora such as internet message boards and movie reviews to train classifiers designed to classify text as positive or negative. The movie reviews corpus of Pang et al. (2002), for example, has become a classic of Sentiment Analysis and is now used as a teaching tool (Bird et al., 2009). Although the widely used MPQA opinion corpus (Wilson and Wiebe, 2005) is manually annotated for multiple attitude types, including "positive and negative arguing," this corpus is newswire text rather than recognizably stancetaking text such as political debate or essays. In itself, this research emphasis on opinion isn't surprising given the field's roots in commercial applications such as the automated classification of online movie and product reviews. However, a consequence of this emphasis on the detection of positive or negative opinion has been a lack of annotated resources dealing with other forms of sentiment such as stance.

One solution to this sparse corpora problem for researchers interested in supervised classification of political perspective or stance has been the use of "self-annotated" corpora. These corpora include congressional floor speeches, where the party affiliation and voting behavior of the speaker serves as the perspective tag (Thomas et al., 2006), the Death Penalty Corpus (Greene, 2007), which contains pro- and anti-death penalty text scraped from websites that self-identify as pro- or anti-death penalty, the Bitter Lemons corpus (Lin et al., 2006), which contains articles dealing with the Israel-Palestine conflict written from an Israeli or Palestinian perspective, and the previously noted Internet Arguments Corpus (Walker et al., 2012b).

The research most similar to the annotation work reported here is Curran and Koprinska (2013). Curran and Koprinska describe the creation of a corpus of *position statements* responding to 7 different topics, with each position statement tagged as *supporting*, *opposing*, or *neutral*. Each position statement originally appeared as a direct quote in a news article related to one of seven manually-defined topics. The articles themselves were extracted from a corpus of newswire using text search software. O'Keefe et al. collected annotations for both position statements lacking any sort of context and position statements considered in the context of text surrounding the quote. The resulting corpus of 2245 sentences was collected using three annotators. The observed agreement score and Fleiss' κ for the no-context set of annotations was .70 and .36, respectively, while observed agreement and Fleiss' κ was .68 and .32, respectively, for in-context annotations.

4.2 Processing the ICLE corpus

As we argued in section 3.7, the discourse context associated with stancetaking language involves a proposition, a subject holding that proposition, and a second subject aligning or disaligning herself with the first subject by arguing *for* or *against* that proposition. We also assumed a scenario where the writer of the essay prompt served as the first subject, the proposition contained in the prompt served as the targeted proposition, and the essay writer served as the second subject aligning or disaligning herself with the proposition. To maintain this scenario, we chose only those essay prompts that contain propositions—these are most likely to elicit essay responses displaying the features of stancetaking language discussed in chapter 3. Additionally, we want to maintain a relatively uniform distribution of topics in our corpus. Thus, topics were included in the set of seven topics only if there were more than 100 essay responses to a topic containing one or more propositions. This rules out prompts that do not direct students to argue *for* or *against* a proposition. Such prompts include those containing *alternative questions* (Karttunen, 1977) which carry the presupposition that one of the alternatives in the question is true (e.g., "Europe: loss of sovereignty or birth of a nation?") and prompts containing *wh*-questions such as "In his novel Animal Farm George Orwell wrote, 'All men are equal: but some are more equal than others' [sic.] How true is this today?" Since ICLE essay files do not include information about the essay topic, the text of each of the seven prompts was used as part of a regular expression pattern search and matched essays were extracted.

Looking at the distribution of topics and stance polarities in Table 4.1, it is clear that we could not maintain a completely uniform distribution of topics and stance polarities. The *Degrees* and *Science* essays give us a somewhat skewed topic distribution toward these two prompts while the *Armies*, *Marx*, and *Prisons* prompts are all skewed heavily toward a single polarity. We find a similarly skewed topic distribution in Somasundaran's (2010) online debates corpus which contains three times as many posts dealing with gay rights (1186) as healthcare (336). For polarity

ESSAY PROMPT	#ESSAYS	%FOR	%AGAINST	%NEITHER
Armies	126	.74	.19	.07
Degrees	279	.51	.45	.04
Feminism	153	.27	.64	.09
Marx	156	.83	.13	.04
Money	103	.49	.42	.09
Prisons	126	.74	.19	.07
Science	377	.12	.86	.02
Total	**1320**	**.45**	**.50**	**.05**

Table 4.1: Distribution of essay responses provided to annotators along with percentages of gold-standard *for*, *against*, and NEITHER tags for each prompt.

classification models trained on feature sets consisting of word presence or frequency, the danger of a skewed topic distribution is that the resulting class probabilities learned by such models will be topic-specific and will not generalize well to other datasets. For example, students arguing for the proposition in the *Marx* prompt, *Television is the opium of the people*, make frequent use of the word *drug*, as in *rulers who use religion as a drug to numb people's awareness* or *television has become such a powerful thing, many of us are addicted to it, like it were a drug*. Yet it is obvious that *drug* is fairly useless as an indicator of *for* stance across other domains such as the *Armies* and *Degrees* prompts. Somasundaran (2010, pg. 191) reports that topic-specific stance features are less likely to occur in lexicon-based feature sets, which make use of stance and sentiment lexicons that are designed to be generalizable across topics. Since our feature sets are also lexicon-based, we hypothesize that the resulting features produced by the procedures described in chapter 5 will also not be topic-specific despite the skewed distribution of topics evident in Table 4.1.

All annotation work was completed using non-expert annotators recruited from online crowdsourcing services. Crowdsourcing services such as Amazon Mechanical Turk (AMT)[1] and Crowdflower (CF)[2] have become popular resources for non-expert annotation of linguistic data for use in diverse NLP applications. These services are a cost-efficient alternative to the often expen-

[1] http://www.mturk.com/mturk

[2] http://www.crowdflower.com/

sive, time-consuming task of expert linguistic annotation. Recent studies of AMT's potential for rapid collection of annotated linguistic data (Snow et al., 2008; Sheng et al., 2008; Callison-Burch, 2009; Callison-Burch and Dredze, 2010) include high interannotator agreement scores between expert and non-expert annotators across a range of linguistic annotation tasks.

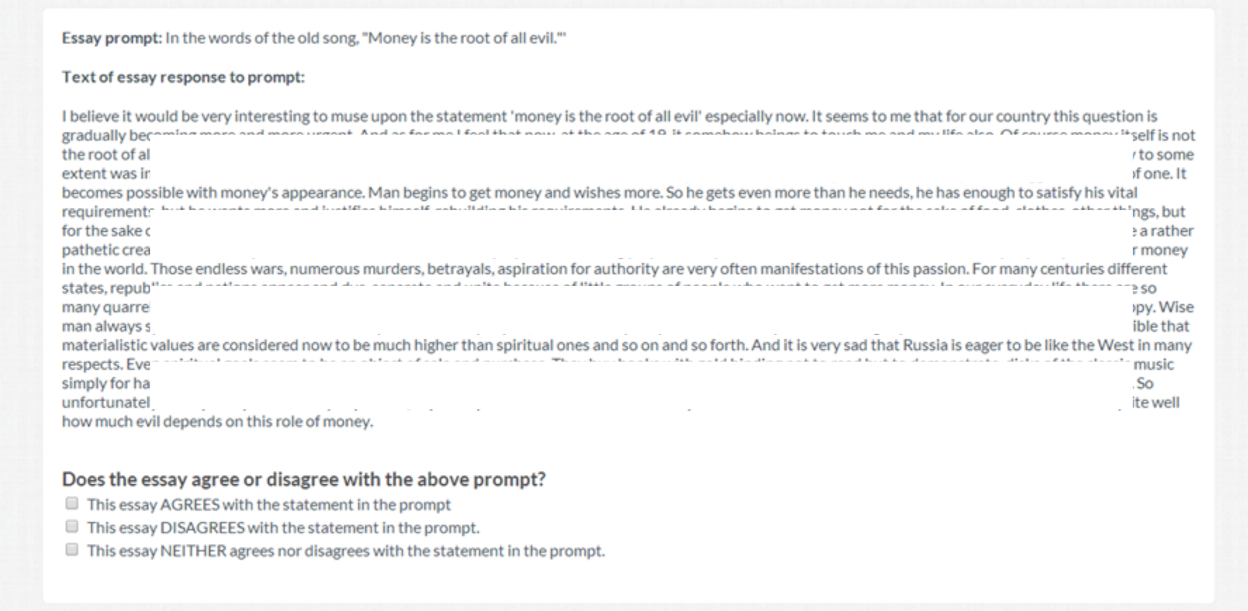
In the area of Sentiment Analysis, AMT has proven to be a reliable source of non-expert annotations for sentiment polarity due to the relative simplicity of the annotation tasks involved. To create a lexicon of emotion-denoting terms, Mohammad et al. (2009) elicited the judgments of five AMT workers for each of $\sim$2000 English terms and reported agreement between at least four of five workers for over 80% of terms. Mellebeek et al. (2010) conducted a document-level annotation study that involved the tagging of Spanish consumer comments for sentiment polarity. Interannotator agreement was calculated between an expert and the majority tag of three unique AMT annotators. Cohen's κ was .72 for expert/majority-AMT-tag agreement.

4.3 Annotation process

All 1320 ICLE essays were posted to CF and five unique annotators were recruited to annotate each essay. For each essay, the essay prompt associated with that essay was included on the same screen. Annotators were asked to read both the essay prompt and the essay in their entirety and to tag each essay as displaying a *for* or *against* stance toward the given prompt. The annotation instructions for this task are given in Appendix A. A screenshot of the CF interface for this task is given in Table 4.1. Gold-standard annotation was performed by the author.

4.3.1 Annotation evaluation

Interannotator agreement was calculated between the gold-standard and the CF-tagged corpus. Annotation tags for the CF-tagged corpus were determined using majority voting and random tie-breaking (Snow et al., 2008). Agreement adjusted for chance was calculated using Cohen's (1960)



Essay prompt: In the words of the old song, "Money is the root of all evil."'

Text of essay response to prompt:

I believe it would be very interesting to muse upon the statement 'money is the root of all evil' especially now. It seems to me that for our country this question is gradually bec [redacted] self is not the root of al [redacted] to some extent was ir [redacted] f one. It becomes possible with money's appearance. Man begins to get money and wishes more. So he gets even more than he needs, he has enough to satisfy his vital requirements [redacted] ngs, but for the sake c [redacted] a rather pathetic crea [redacted] r money in the world. Those endless wars, numerous murders, betrayals, aspiration for authority are very often manifestations of this passion. For many centuries different states, repub [redacted] e so many quarre [redacted] py. Wise man always s [redacted] ible that materialistic values are considered now to be much higher than spiritual ones and so on and so forth. And it is very sad that Russia is eager to be like the West in many respects. Eve [redacted] music simply for ha [redacted] . So unfortunatel [redacted] ite well how much evil depends on this role of money.

Does the essay agree or disagree with the above prompt?

☐ This essay AGREES with the statement in the prompt
☐ This essay DISAGREES with the statement in the prompt.
☐ This essay NEITHER agrees nor disagrees with the statement in the prompt.

Figure 4.1: Screenshot of the CF interface for the essay-level stance annotation task.

κ, which is

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)},$$

where $P(a)$ is the observed percentage of agreement between two annotators, and $P(e)$ is the agreement expected to occur by chance as might happen if the annotators simply click on the *for*-arguing, *against*-arguing, or *neither*-arguing buttons in a random manner. The resulting score, then, can be interpreted as observed agreement after adjusting for any agreement that occurs by chance. Both observed interannotator agreement and κ are given in Table 4.2.

Cohen's κ was .68, which Landis and Koch (1977) interpret as "substantial." Note that this compares favorably with the .72 score of Mellebeek et al. To get a sense of how this result compares with similar work dealing with three-category document-level annotation of sentiment polarity, we can examine the agreement scores reported in O'Hare et al. (2009) for a financial blog sentiment annotation task. The corpus of O'Hare et al. (2009) is similar to ours in size (979 documents) and in writing style, as financial blogs tend to avoid the highly emotional language of movie and product reviews. They report a Cohen's κ of .71 for a three-category (*positive*, *nega-*

Prompt topic	Observed	Cohen's κ
Armies	0.91	0.79
Degrees	0.76	0.61
Feminism	0.79	0.59
Marx	0.87	0.57
Money	0.78	0.63
Prison	0.82	0.59
Science	0.83	0.51
All prompt topics	**0.82**	**0.68**

Table 4.2: Observed agreement and Cohen's κ score for essay-level stance annotation across all 7 ICLE prompts.

		Crowdflower			
		for	against	Neither	Total
Gold	for	**497**	62	24	583
	against	39	**563**	23	625
	Neither	56	34	**22**	112
	Total	592	659	69	**1320**

Table 4.3: Contingency table for the essay-level stance annotation task.

tive, *neutral*) annotation task, which compares favorably to our score. Thus, while annotators find stance annotation somewhat more difficult than the similar task of opinion annotation, the substantial agreement score achieved for our task indicates that even untrained annotators have very clear intuitions regarding document-level stance polarity.

4.3.2 Error analysis

Examining instances of annotator disagreement, we find that cases of disagreement involving the existence of a stance polarity (i.e., for a given essay, the majority of annotators marked an essay as *neither for nor against* while the gold-standard tag was either *for* or *against*) usually involve responses to prompts containing propositions with multiple premises such as the *Prisons* and *Degrees* prompts. When a writer argues *for* one or more of the premises of an argument contained in the prompt text, but argues *against* the conclusion implied by those premises, annotators seem to consider this sufficient reason to tag the essay as *neither*. For example, in the concluding sentences of an essay responding to the *Prisons* prompt in (35), we find the writer arguing for certain premises of the prompt statement while arguing against others (stances taken toward propositions in the prompt have been boldfaced and the polarity of these stances have been given in brackets). The essay from which (35) was excerpted was tagged *neither* by the majority of CF annotators.

(35) As it must have become clear to you, **I am not really convinced of the efficiency of the rehabilitation system** [*against*], although I must recognize that it can work in some cases, and that it is a praiseworthy alternative. **I agree that our prison system is outdated** [*for*], or not appropriate, but **rehabilitation is far from being the solution to all the problems.** [*against*]

In other cases of disagreement involving the *neither* tag, annotators tended to tag any essay that takes a stance toward the premise of a proposition, but takes no explicit stance taken toward the conclusion of that proposition, as *neither*. For example, the *Degrees* prompt contains the premise *Most university degrees are theoretical and do not prepare students for the real world* followed by the conclusion, *They are therefore of very little value*. Sentence (36a) argues *for* this premise while sentence (36b) argues *against*. Yet the essays from which (36a) and (36b) were taken were both tagged as *neither* by annotators since both essays lack a *for* or *against* argument regarding the conclusion, *They are therefore of little value*.

(36) a. To my mind it is not enough for an university to have mainly theoretical classes and not to prepare practically those who wish to get a university diploma.

b. If you are well-educated theoretically you have all chances to achieve success in real world.

In such cases, the usual inference that the validity of *Most university degrees are theoretical and do not prepare students for the real world* implies the validity of *They are therefore of very little value*, does not seem to agree with annotator intuitions. The existence of sentences arguing *for* (*against*) one or more of the premises contained in the prompt is not, apparently, sufficient reason to tag an essay as *for* or *against*.

In cases of disagreement regarding an essay's *for* or *against* stance, the issue often does not involve varying intuitions regarding logical inference; instead, the ambiguous use of rhetorical devices is the main culprit. (37a) is the lead sentence of an essay that is disaligned with the *Feminism* prompt. Yet the underlying assertion of such "*wh*-reversed polarity questions" (Koshik, 2005), reconstructed by the reader as the declarative stripped of its *wh*-question material in (37b), seems to be aligned with the prompt.

(37) a. What is feminism and **why on earth has it done so much harm to women**?
b. It has done so much harm to women.

Based only on (37a), we might be tempted to tag the essay from which this sentence was extracted as aligned with the *Feminism* prompt since it seems to endorse the claim that feminism has done harm to women. When we read further, however, clear indications of the writer's disaligned stance emerge, exemplified by the section in (38).

(38) Women of today would not have choice whether to work or to be a houswife but for the feminists. They proved that women are not worse than men. Feminists were one of the motives of the world progress inspite of the fact that they were sometimes mistaken.

In other cases of *for/against* disagreement, a sentence explicitly aligned (disaligned) with the prompt functions as part of a broader argument which in turn takes an opposite polarity stance to that of the sentence. Writing in response to the *Marx* prompt, the writer of (39) aligns herself with the proposition *television is the opium of the people* by arguing that the media encourages narcissism.

(39) All the mass media convince us daily that every man, every member of society is a unique individual, that human intellect is almost all-powerful, that it has boundless opportunities and that a man should enjoy himself since human life is short.

In the next sentence, given in (40), the writer takes an explicit *against* stance toward the main proposition.

(40) I think we can come to the only conclusion that not television itself is the opium, but the aforesaid idea.

Taken together, (39) and (40) argue that television is simply a facilitator of the ideas given in (39), yet the appearance of (40) appears to be enough to convince annotators that the writer is arguing against the proposition in the *Marx* prompt.

4.3.3 Discussion

We have seen that annotator disagreement regarding the polarity or existence of stance at the document level can arise as the result of differing intuitions regarding logical validity. This problem also arises in the argument annotation work reported in (Palau and Moens, 2009). Palau and Moens (2009) describe the annotation process involved in the creation of a "legal argumentation" corpus based on Walton's (2008) influential schemes of informal logical argumentation. Working with a corpus of documents consisting of legal cases from the European Court of Human Rights, annotators were asked to tag sections of text according to Walton's schemes. Palau and Moens (2009) report that the most consistent source of disagreement between annotators involved the assessment of an argumentative clause as a premise or conclusion of a larger argument. Thus we find that, in general, annotator disagreement for tasks involving either the explicit or implicit recognition of valid argument structure tends to be the result of a basic conflict between intuition- and rule-based conceptions of logical validity.

Chapter 5

A feature scheme for essay-level stance classification

In this chapter, we present our feature scheme for essay-level stance classification. We describe the process of engineering two different feature sets for use in the classification experiments reported in chapter 6. Our approach to the engineering of classification features is motivated by the linguistic observations of chapter 3 and draws on similar work in the areas of opinion mining and stance classification.

5.1 Related work

We first describe a feature set that captures the relationship between a stancetaking word identified in a given clause and the proposition targeted by that word in the immediate or embedded S-clause. In the literature, there are only a handful of unsupervised approaches to the task of identifying the targets of evaluative language. The OPINE system of Popescu and Etzioni (2005) is an example of a fine-grained approach to the extraction of $\langle opinion, aspect \rangle$ tuples where an *opinion* is some sentiment-bearing phrase and an *aspect* is some aspect of a global opinion target. Popescu and Etzioni experimented with a corpus of product reviews which contain global opinion targets such as scanners and phones. Aspects of a scanner could be *scan quality*, *battery life*, *cover*, etc. To identify these aspects and the opinion phrases targeting them, candidate opinion phrases were extracted using a set of hand-crafted rules. The polarity of each candidate is calculated as the difference between the web-based Pointwise Mutual Information (PMI) score of that phrase and the score of manually selected positive (negative) keywords, such as *excellent* and *awful*. The aspects targeted by the opinions were identified by first extracting all noun phrases in the neighborhood of the identified opinion word and identifying any meronymic (part-of) relationships between the noun phrase and the product under review. Meronymic relationships are determined using the PMI score between the potential part or property and phrases such as *of scanner*, *scanner has*, etc.

Other more fine-grained, unsupervised approaches to the identification of opinion targets include the system of Kim and Hovy (2006), who used FrameNet (Fillmore et al., 2003) relationships to determine both the opinion source and target of a given opinion word, and Qiu et al. (2011) who used manually identified relationships between opinion words and their targets in dependency-parsed text to iteratively identify additional relationships and further examples of opinion/target pairs.

In the area of stance classification, there are two examples of stance-target identification: Somasundaran and Wiebe (2010) and Anand et al. (2011). As Somasundaran (Somasundaran, 2010, pg. 160) observes, the fact that the targets of opinion tend to be nominal while those of stancetaking tend to be propositional means that the approaches to opinion target identification mentioned above are not applicable to the task of stance target identification. The alternative approach of Somasundaran and Wiebe is coarse-grained: Given a sentence containing a stancetaking word, all content words in that sentence are considered part of the proposition target and are tagged as such. This results in a very noisy stance-target feature set, as discussed further in section 5.2.1. The approach of Anand et al. is also coarse-grained: All sentences are dependency-parsed and the heads of the resulting tuples are considered stance targets.

In section 5.3, we present a feature set that captures the relationship between the stancetaking language of a given ICLE essay and the language of the prompt to which the essay is responding. Our approach involves relating the content words of a particular essay's prompt to any content words located in the immediate or embedded clause containing an identified stance word. Of course, more of often than not, an essay writer will deal with some aspect of the prompt by using a different, though semantically related word. In an essay response to the *Money* prompt, for example, we may find the writer using a word that is lexically related to the prompt word *evil*, such as the near-synonym *demonic*, or a word with some cultural association with *evil* such as the literary character *Raskolnikov*. In either case, the writer is using a word bearing a relationship to the word *evil* and is thus dealing with some aspect of the prompt.

We view the task of linking prompt words (or words semantically related to prompt words) to words in an essay as analogous to the task in aspect-based Sentiment Analysis (see section 2.4.1) of identifying aspects of the topic of an opinion. Just as an *interface* is an aspect of *iPhone* and *foreign policy* is an aspect of *Obama*, *demonic* and *Raskolnikov* are aspects of the text of the essay prompt which serves as the global target of the essay. A key component of aspect-based Sentiment Analysis is the choice of a semantic similarity metric. Historically, there have been two approaches in NLP to the task of capturing the semantic similarity between words, one utilizing statsitical properties of large corpora and the other utilizing pre-existing knowledge resources. Popular corpus-based approach include vector similarity metrics such as Latent Semantic Analysis (Landauer et al., 1998), which exploits the fact that semantically related words tend to occur in very similar contexts, and information-theoretic metrics such as Pointwise Mutual Information.

Pre-existing knowledge sources include lexical thesauri such as WordNet (Miller, 1995; Fellbaum, 1998) and ontologies such as ConceptNet (Liu and Singh, 2004). Recently, the problem of discovering non-lexical semantic relationships across varied domains, such as the relationship between *evil* and *Raskolnikov*, was addressed in work exploiting the vast repository of semi-structured knowledge found on the internet. The Normalized Google Distance metric of Cilibrasi and Vitanyi (2007) is one such measure but is unreliable because the hit counts utilized in this formula change from day to day (Funahashi and Yamana, 2010). Additionally, Google's search API can be prohibitively expensive for researchers. We use an alternative semantic similarity metric, the Wikipedia Link-based Measure of Witten and Milne (2008), to collect more reliable semantic similarity scores across unrestricted domains.

5.2 Stance-attitude features

In this section, the descriptive generalizations discussed in chapter 3 motivate the creation of a set of classification features that capture the relationship between a stancetaking word and its proposition. During this process, we draw on the following three linguistic generalizations:

1. At the semantic level, the targets of stancetaking language are propositions (section 3.5).

2. Syntactically, propositions are realized as clauses and are located in the embedded S-clause of a stance word taking a *that*-clause object or in the immediate S-clause of a given stance word (section 3.6.1).

3. At the lexical level, stancetaking can be reduced to a small class of words that includes evidentials (*possibly, should, indicate*) and metadiscourse (*in addition, moreover*) (sections 3.6.1 and 3.7). Additionally, Pollyanna effects are evident in stancetaking lexis (section 3.6.2). One consequence of these effects is that *against* stance is usually expressed by simply negating *for*-stanced expressions.

In order to capture the first and second of these generalizations as classification features, we assume that any attitudinal language found in the immediate or embedded S-clause of a stance expression can be considered a proxy for the full proposition targeted by that expression. This strategy is motivated by Martin and White's notion that stancetaking involves a speaker/writer aligning (disaligning) herself with an "attitudinal assessment" (Martin and White, 2005, pg. 95).

We also introduce a formalism to capture this relationship, the *stance-attitude profile*. To identify the clausal relationship between a stance expression and the attitudinal language targeted by that expression we first parse all sentences using a standard syntactic parser trained on the Penn Treebank.

To capture the third generalization, we use two knowledge resources containing stancetaking and attitudinal words. The first resource is a stance lexicon comprised of stancetaking words from the lexicon created in Somasundaran and Wiebe (2010) and a selection of words from Hyland's (2005) list of metadiscourse markers. The second is the list of attitudinal words contained in the MPQA subjectivity lexicon (Wilson and Wiebe, 2005). These resources will be used to identify stancetaking and attitudinal words in a given ICLE essay. In order to capture both short- and long-distance negation of *for*-stanced words we dependency-parse all sentences using the Stanford parser and append the tag "not" to all *for*-stanced words modified by a negator in the parse.

5.2.1 Linking expressions of stance with their proposition targets

In the following sections we present our approach to identifying markers of stance in an ICLE essay and locating the propositions targeted by those markers. Our first task is devise a way of succinctly representing the stance/proposition relationship as a classification feature. We cannot simply extract the complete clausal proposition as a stance target since this would give us a very poor, non-generalizable classification feature. We also cannot adopt the approach taken in Popescu and Etzioni (2005) who identified any nominal elements in the vicinity of opinion words as the targets of those words. In our case, the targets of interest are clausal, not nominal. Our strategy is to find elements in the clause that can serve as proxies for full propositions. This approach is partly inspired by Somasundaran (2010) who considered all content words in a sentence containing a dominant positive (negative) arguing stance as proxies for the proposition targeted by that stance. Tags indicating the sentence's dominant stance were then appended to each content word. When one applies this strategy to a sentence from an essay responding to the *Marx* prompt, given in

(41a), the result is the transformed sentence given in (41b). Each stemmed content word receives an *ap* (=arguing positive) tag since the sentence contains two positive arguing stance words (*could, would*).

(41) a. If only Karl Marx could guess about the possibility of inventing television, he would at first rejoice at it: television provides all those involved in political and social life of masses with a priceless means of great influence on people's minds.
b. `<ap-only, ap-karl, ap-marx, ap-could, ap-guess, ap-possibl, ap-invent, ap-televis, ap-first, ap-rejoic, ap-televis, ap-provid, ap-involv, ap-polit, ap-social, ap-life, ap-mass, ap-priceless, ap-mean, ap-great, ap-influenc, ap-people>`

Examining the transformed sentence given in (41b), it is clear that using content words as proxies for propositions results in a very noisy feature set. First, by indiscriminately tagging all content words, it is not clear which segments of the sentence are meant to include the propositions targeted by *could* and *would*. Second, many of the resulting features are topic-specific and will likely not generalize well beyond the *Marx* section of the corpus. These include such features as *karl_ap, marx_ap, televis_ap* and *polit_ap*.

Our alternative proposal is to consider any attitudinal words in the immediate or embedded clause relative to the clause containing a stance word as proxies for full propositions. This approach results in features that succinctly capture the relationship between a given stancetaking word and its proposition and has the additional advantage of being non-topic specific. This strategy was inspired by Martin and White's (2005) conception of speaker/writer alignment (disalignment) as the act of agreeing (disagreeing) with a given "attitudinal assessment" (Martin and White, 2005, pg. 137), that is, with a proposition containing attitudinal language. While it is not always the case that a proposition will contain attitudinal language, each of the ICLE prompts contains at least one such proposition, as shown in (42). All attitudinal language is identified by matching a stemmed version of each word in the prompt to a stemmed version of the MPQA subjectivity lexicon (Wilson and Wiebe, 2005). We also assume in these examples that negators (*no, not*) and downtoners (*little*) reverse the polarity of the attitude word that they modify. In the *Science* prompt, for example, given in (42g), *no* is used to reverse the polarity of the positive words *imagination* and *dreaming*.

(42) a. There is **no value [*negative*]** in a system of military service.

b. Most university degrees are theoretical and do **not prepare [*negative*]** students for the **real [*positive*]** world. They are therefore of very **little value [*negative*]**.

c. Feminists have done more **harm [*negative*]** to the cause of women than **good [*positive*]**.

d. If he was alive at the end of the 20th century, he would **replace [*negative*]** religion with television.

e. In the words of the old song, "Money is the root of all **evil [*negative*]**."

f. The **prison [*negative*]** system is **outdated [*negative*]**. No **civilized [*negative*]** society should **punish [*negative*]** its **criminals [*negative*]**: it should **rehabilitate [*positive*]** them.

g. Some people say that in our **modern [*positive*]** world, **dominated [*negative*]** by science, technology, and industrialization, there is **no** longer a place for **dreaming [*negative*]** and **imagination [*negative*]**.

In our discussion of the discourse relationship between stancetaking language in an ICLE essay and the language of the prompt to which it is responding (section 3.7), we saw that the clearest indications of a writer's stance toward the proposition contained in the prompt occur in those sentences that selectively reproduce the language of the prompt and modify that language using markers of stance. We can capture this in our feature representations by reducing the propositions in both the prompt and the essay response to any attitude words found in that proposition. To illustrate this idea, consider the *Prisons* prompt, separated into the two propositions (43a) and (43b). As in (42) we can isolate any attitudinal lexis by referencing the MPQA subjectivity lexicon. Additionally, to identify stancetaking lexis, we can again reference the lexicon of stancetaking language constructed by Somasundaran (2010) and Somasundaran and Wiebe (2010).

(43) **Prompt:** The prison system is outdated. No civilized society should punish its criminals.

a. No civilized society should punish its criminals.

Positive stancetaking lexis: n/a
Negative stancetaking lexis: no should

Positive attitudinal lexis: n/a
Negative attitudinal lexis: punish, criminal

b. It should rehabilitate them.

Positive stancetaking lexis: should
Negative stancetaking lexis: n/a

Positive attitudinal lexis: rehabilitate
Negative attitudinal lexis: n/a

We first note that (43) contains two acts of stancetaking, one arguing against incarceration as punishment (43a) and another arguing for rehabilitation as an alternative (43b). Rhetorical relationships of antithesis (Mann and Thompson, 1988, pg. 253), with the second argument serving as a counterpoint to the first, are not uncommon in stancetaking text. Maintaining our assumption that attitude words are proxies for propositions, all instances of stancetaking in (43) can be viewed as simple lexical relations with a stance word taking an attitude word as a target (in the examples given in this section, we restrict our attention to stance-attitude pairs occurring in the immediate clause). We can also represent each stance-attitude tuple as combinations of *S+/-* and *A+/-* where *S* is *stance* and *A* is *attitude*. The polarities +/- of *S* and *A* map to *for* and *against* in the case of *S*, and *positive* and *negative* in the case of *A*. Using this representation, (43a) can be represented as (44a) and (43b) can be represented as (44b).

(44) a. (no should, punish) = *S-A-*
(no should, criminals) = *S-A-*

b. (should, rehabilitate) = *S*+*A*+

This representation captures the intuition that, in (43a), the prompt author is arguing *against* something negative (*No civilized society should punish its criminals*) and, in (43b), is arguing for something positive as an alternative (*it should rehabilitate them*). Responding to this prompt, the writer of (45) aligns herself with the prompt author regarding her stance on prison reform.

(45) Society would benefit more from rehabilitating its criminals.

Positive stancetaking lexis: would
Negative stancetaking lexis: n/a

Positive attitudinal lexis: benefit, rehabilitating
Negative attitudinal lexis: criminals

(45), which takes the proposition *it should rehabilitate them* as its target, can also be reduced to a $\langle stance, attitude \rangle$ representation, as in (46).

(46) (would, benefit) = *S*+*A*+
(would, rehabilitating) = *S*+*A*+
(would, criminals) = *S*+*A*-

It is also possible to align oneself with the prompt author by targeting both of the propositions in the prompt, as in (47).

(47) Criminals must not be punished but they should be rehabilitated for the benefit of society.
Positive stancetaking lexis: should
Negative stancetaking lexis: must not

Positive attitudinal lexis: rehabilitated, benefit
Negative attitudinal lexis: punished

The $\langle stance, attitude \rangle$ tuple representation of (47) is (48).

(48) (must not, punished) = *S-A-*
(should, rehabilitated) = *S+A+*
(should, benefit) = *S+A+*

How do essay writers *disalign* themselves with one or more of the propositions contained in a prompt statement? Consider example (43a). Maintaining our representation of stance-attitude pairs as two-membered tuples, each member taking one of two polarity values, there are two possible ways of disaligning oneself with the proposition *No civilized society should punish its criminals*. One can reverse the polarity of the attitudinal language in the embedded clause *punish its criminals* by using a positive polarity antonym of *punish* such as *help*, while maintaining the polarity of the stance marker *no should* —the result might be *No society should help its criminals*. The alternative strategy, used by the author of (49), is to maintain the negative polarity of the attitudinal language in the proposition while reversing the polarity of the stancetaking language.

(49) One must preserve a system in which criminals are punished for their prejudicial deeds.
Positive stancetaking lexis: must
Negative stancetaking lexis: n/a

Positive attitudinal lexis: n/a
Negative stancetaking lexis: criminals, punished, prejudicial

The stance-attitude tuples for (49) are given in (50).

(50) (must, criminals) = *S+A-*
(must, punished) = *S+A-*
(must, prejudicial) = *S+A-*

5.2.2 Stance-attitude profiles

Examples (44-50) suggest that the lexical relationships between stance and attitude can be represented as a 2 × 2 contingency table, as given in Table 5.1. Table 5.1 presents all possible combinations of stance and attitude polarities. We call the entry in each quadrant a *stance-attitude* *profile*. Importantly, Table 5.1 allows us to model the student's act of alignment (disalignment) with the prompt's stance as the selection of a quadrant along the diagonal of the quadrant corresponding to that stance (in the case of alignment) or off the diagonal of the quadrant corresponding to that stance (in the case of disalignment). For example, prompt (43a) has the quadrant value *S*-*A*-. A writer arguing in alignment with (43a) would then choose one of the quadrants along the diagonal of this quadrant—i.e., *S*-*A*- or *S*+*A*+—as occurs in (45-48). By contrast, a writer arguing in disalignment with (43a) would choose a quadrant *off* the diagonal of *S*-*A*-, either *S*-*A*+ or *S*+*A*-. The writer of (49) chooses the latter.

	Attitude	
Stance	+	−
$+$	$S+A+$	$S+A-$
$-$	$S-A+$	$S-A-$

Table 5.1: Stance-attitude combinations as a 2×2 contingency table

5.2.3 A procedure for extracting stance-attitude features

The notion of stance-attitude profiles described in the last section serves as a template for the classification features described in this section. When constructing our algorithm for stance-attitude profile extraction, our main goal is a procedure that is sufficiently general to capture many of the lexico-syntactic phenomena we discussed in chapter 2. It is first necessary to capture the role of negation in reversing the polarity of stance words. As discussed in section 3.6.2, speakers/writers tend to express an *against* stance by negating positively stanced language. At the same time, we would like to capture the negation-triggered reversal of *for* stance polarity (e.g., in *not deny*, *not* reverses the polarity of *deny* from *against* to *for*). Consider the co-occurrence patterns in (51).

(51) a. I do **not think** feminists have done more harm to the cause of women than good but I think male chauvinists have done it.

b. I do **not** really **believe** that reality has a say in anything that concerns our dreams because somehow by definition dreams and imagination are antidotes to reality.

c. That life can exist without money, I do**n't think** is possible.

d. **Not** so long ago I heard a suggestion that if Marx was alive today he **would** replace religion with television.

In the case of (51a), modification by negation occurs locally. However, modification by negation can also occur across other modifiers as in (51b). In (51c), the extraposed stance marker *think* is negated and must in turn take proper scope over the attitude word *life*. Thus, we must capture both modification of *think* by *n't* and modification of *life* by the extraposed *n't think*. Finally, (51d) exemplifies the problem with a naïve, linear search approach to identifying the targets of negators. The negator modifying the adverbial group *so long ago* should not reverse the polarity of *would* in the lowermost clause.

Our second task involves capturing the clausal relationship between stance words and the propositions targeted by those words—in our case, attitudinal words serve as proxies of the targeted propositions. As we showed in section 3.6.1, stancetaking words target propositions in the immediate S-clause (or in the embedded S-clause in the case of stance words selecting for *that* clause objects). This also holds for doubly embedded stance words. Thus, given a stance word taking a *that*-clause object, our algorithm should look for attitudinal lexis in the embedded S-clause; in all other cases it should look in the immediate clause. Consider the examples in (52). Markers of stance are boldfaced and those of attitude are underlined.

(52) a. The progress of technology often **seems** to have the dangerous effect of cutting down on humans' imaginative and creative powers, but that does **not mean**, however, that they are to perish at a certain point in the future.

b. I do **not think** feminists have done more harm to the cause of women than good but I **think** male chauvinists have done it.

c. I **believe** that if our ancestors have not used their imagination in the past we **would** have still be dreaming of any development in these spheres.

In (52a) the positively stanced *seems* modifies the attitudinal lexis in the adjacent clause *dangerous*, *cutting, imaginative,* and *dreaming*, but not the attitude word *perish* in the lowermost clause which is modified by *not mean*. Similarly, in both (52b) and (52c), the stancetaking language of the topmost clause must be constrained to modify only material in the adjacent clause.

As a first step, we assign to each sentence in the ICLE corpus a structural representation that can capture both short- and long-distance negation. This representation should also allow us to capture idiosyncratic constructions involving extraposed stancetaking lexis such as example (51c). All extraposed stancetaking words are boldfaced and the attitude words over which they must take scope are underlined.

(53) a. No doubt about it— the world has become more cynical and everything or so it **seems**.

b. An army that consists of professional soldiers only offers more opportunities for women to make a career for themselves in the army I **think**.

c. This is a hard and pleasant work I **believe**.

Traditionally, the structural representation of choice in NLP has been the phrase-structure grammar. Yet, for many information retrieval and NLP tasks that involve the recognition of lexical relationships, such as entity-entity relationship extraction and entity-attribute extraction (Sarawagi, 2008, reviewed in), dependency grammars have become a popular alternative to phrase-structure grammars. The latter contain layers of syntactic complexity that obscure the relationships that are the focus of such tasks. Briefly described, structure in a dependency grammar is determined by the relationship between a head word and its dependent with the clause-level finite verb serving as the root of the resulting dependency tree. Sentiment Analysis research utilizing dependency tree representations includes Matsumoto et al. (2005), Wilson et al. (2009), and Nakagawa et al.

(2010).

The dependency representation used here is based on the Stanford Dependency (SD) grammar of De Marneffe and Manning (2008), which contains 56 head-dependent relationships. The SD grammar has been implemented as a publicly available parser which we have made use of during the feature representation process reported here. The relevant sections of the SD parse of the *Feminism* prompt can be drawn as the directed graph in Figure 5.1, with grammatical relationships as edge labels and the direction of the edge ordered as head to dependent.

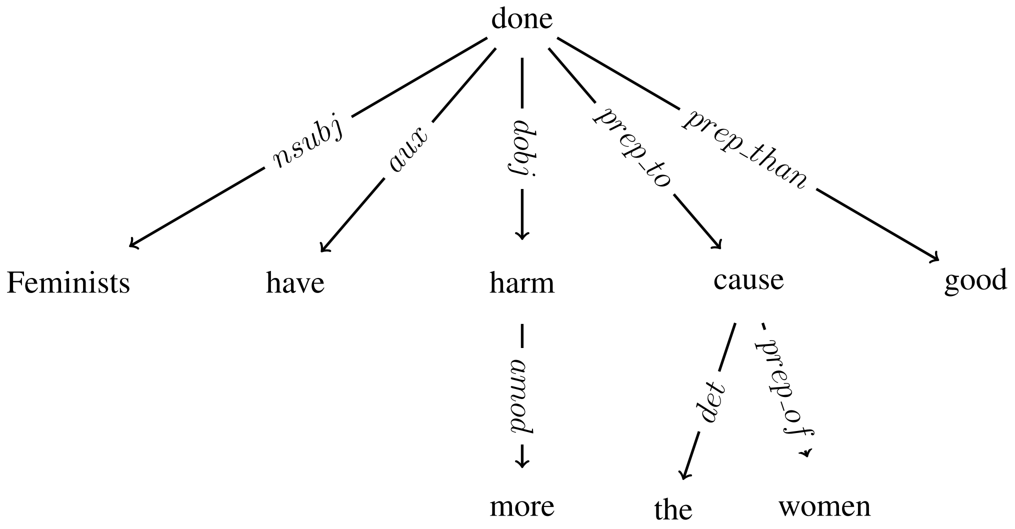


Figure 5.1: Stanford dependency parse of the *Feminism* prompt.

In this example, the target stance-attitude tuples are both $\langle done, harm\rangle$ and $\langle done, good\rangle$. One consequence of the *more-X-than-Y* comparative antonymy construction is that some measure of Y is maintained, i.e., feminism has done some good.

5.2.4 Part-of-speech-generalized dependency sub-trees

In this section we introduce the feature representation scheme we will use in the experiments reported in chapter 6: the part-of-speech (POS) generalized stance-attitude dependency sub-tree.

This feature captures the stance-attitude profile representation described in section 5.2.2. Transforming the language of an essay into a POS-generalized stance-attitude sub-tree feature representation involves 3 basic steps.

- **Step 1**. Each sentence in a given essay is dependency parsed using the Stanford parser.
- **Step 2**. We apply an algorithm that locates any stancetaking and attitudinal words in the parsed sentence and checks to see if the two words meet the clausal constraints described in section 3.6.1. It then finds the shortest path between those words. The algorithm works with an undirected version of a dependency tree, traverses each branch of that tree, starting at a node containing a stancetaking marker and checks if an immediately neighboring node contains a negator. It then traverses the tree in breadth-first fashion by visiting neighboring nodes until it finds a node containing an attitude word. Finally, it checks a phrase-structure parse of the sentence to see if clausal restrictions have been maintained between the two identified words.
- **Step 3**. The mid-nodes between the stance node and opinion node of all sub-trees returned by Step 2 are then POS-generalized.

The resulting feature is a POS-generalized sub-tree with a stance word as a head and an attitude word as a tail. The algorithm used in Step 2 is given as Figure 5.2.

The sub-tree extraction algorithm relies on two external functions, one for breadth-first-search (`BFS`), and another (`clauseRest`) that checks that the extracted stance and attitude words meet the clausal restrictions discussed in section 3.6.1. Since dependency parses, by definition, do not contain the hierarchical structure we require for our clausal constraints check, the original sentence is phrase-structure parsed into Penn Treebank format using the Stanford parser.

The stance-attitude dependency sub-trees extracted using the algorithm given in Figure 5.2 can be generalized by "backing-off" each word in the mid-path of the sub-tree to its POS. This

```
function BFS(graph, node1, node2)                          ▷ Function for breadth-first-search
function CLAUSEREST(word1, word2)                          ▷ Function for clause restrictions check

$subTree \leftarrow None$
$stance \leftarrow None$
$attitude \leftarrow None$
$neg \leftarrow False$

for $node$ in $sentenceGraph$ do
    for $immediateNeighbor$ in $sentenceGraph$ do
        if $node$ in $stanceWords$ then
            $stance \leftarrow node$
            if $immediateNeighbor$=$neg$ then              ▷ Check if $stance$ is negated
                $neg \leftarrow True$
        else if $node$ in $attitudeWords$ then
            $attitude \leftarrow node$

if $stance \neq None$ then
    if $attitude \neq None$ then
        if clauseRest$(stance, attitude) = True$ then      ▷ Check clausal restrictions
            $subTree \leftarrow$ BFS$(sentenceGraph, stance, attitude)$   ▷ Get shortest path
if $neg$=$True$ then                                       ▷ If $neg$=$True$ append "not" to $subTree$
    $subTree \leftarrow$ not+$subTree$
return $subTree$
```

Figure 5.2: An algorithm designed to extract POS-generalized stance-attitude subtrees given a dependency parsed sentence graph, a stance lexicon, and an attitude lexicon.

process of feature generalization using POS backoff is inspired by a similar approach used in Joshi and Penstein-Rosé (2009). Joshi and Penstein-Rosé (2009) experimented with POS-backoff in the context of an opinion mining task. They reported a significant increase in accuracy when this technique was applied to a bag-of-dependency-tuples feature set. Dependency tuples such as $\langle overwhelming, evidence \rangle$ were generalized to tuple members' POS to create tuples such as $\langle JJ, evidence \rangle$ and $\langle overwhelming, NN \rangle$, where *JJ* and *NN* are Penn Treebank tags for adjective and noun, respectively.

In our case, we are interested in capturing sub-tree patterns extracted by the sub-tree extraction algorithm that are highly discriminative of *for* and *against* stance. Thus, we retain the head and tail words of our extracted subtrees—these correspond to stance and attitude words, respectively—and POS-generalize all words in the mid-path. This approach is motivated by the observation that many expressions involving the interaction of stance and attitude in the ICLE corpus fall into predictable lexico-syntactic patterns characteristic of "formulaic language" (Wray, 2005). Formulaic language consists of fixed or semi-fixed lexico syntactic chunks such as the pattern *a-matter-of-V-ing* (Wray, 2002: 25), which produces *a matter of developing skills, a matter of learning, a matter of becoming able to*, etc. The *can V true* pattern, where *can* is a *for* stance word and *true* is a positive attitude word, is often associated with positively stanced language in the ICLE corpus. Examples of POS-generalized subtrees containing this pattern are given in Table 5.3. All POS tags, as returned by the Stanford POS-tagger (Klein and Manning, 2003), are appended to the word found in each node. The original Penn Treebank tags returned by the POS-tagger have been generalized to their basic grammatical categories—e.g., *VBD, VBG*, and *VBZ* are all simply V. Successful search paths are indicated by dotted arrows and boxed nodes indicate a match for stance or attitude words. The extracted sub-trees are given in column 2, while the stance-attitude profile corresponding to stance and attitude word combinations are given in column 3.

In Figure 5.4, we provide additional illustrations of the algorithm given above. This figure shows the POS-generalized subtrees extracted from (52a), which contains an example of extra-

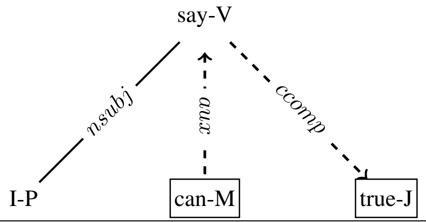
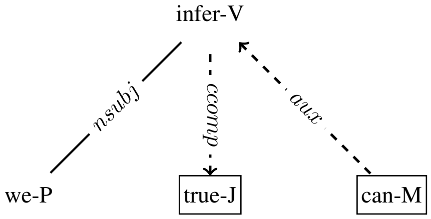
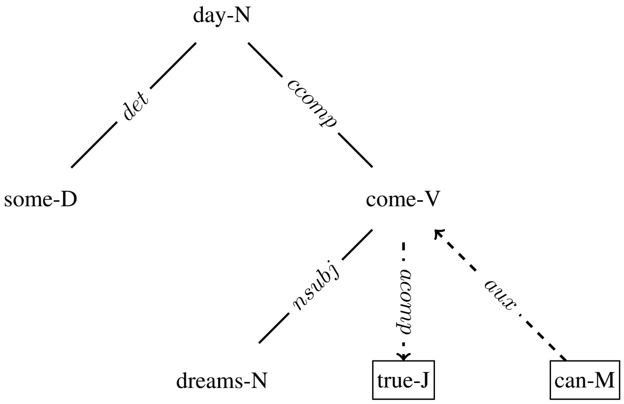
DEPENDENCY TREE TRAVERSAL	POS-GENERALIZED SUB-TREE	STANCE-ATTITUDE PROFILE
“I can only say that this statement is completely **true**”  	*can → V → true*	$S+A+$
“So we can infer that the statement is very **true**.”  	*can → V → true*	$S+A+$
“Some day our dreams can come **true**.”  day-N det ccomp some-D come-V nsubj acomp aux dreams-N true-J can-M	*can → V → true*	$S+A+$

Figure 5.3: POS-generalized sub-tree extraction involving the *can-V-true* pattern. Boxed nodes indicate words matched to stance and attitude lexicons. Sub-tree path traversal is indicated by a dotted line.

posed stancetaking, and (53a), which contains an example of negated *for*-stanced language.

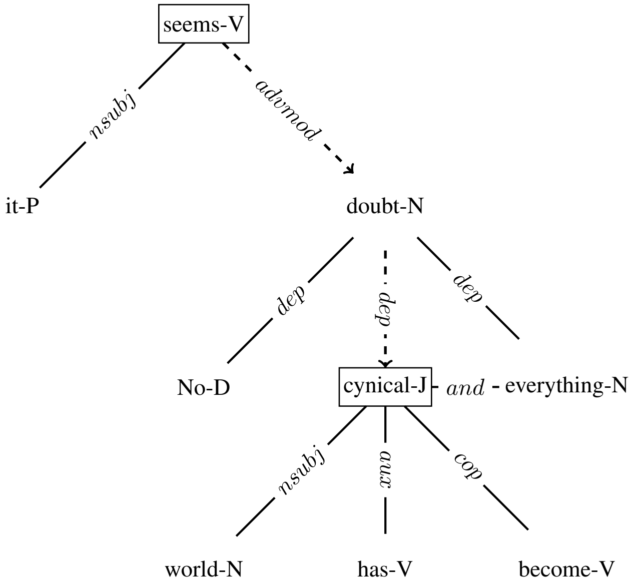
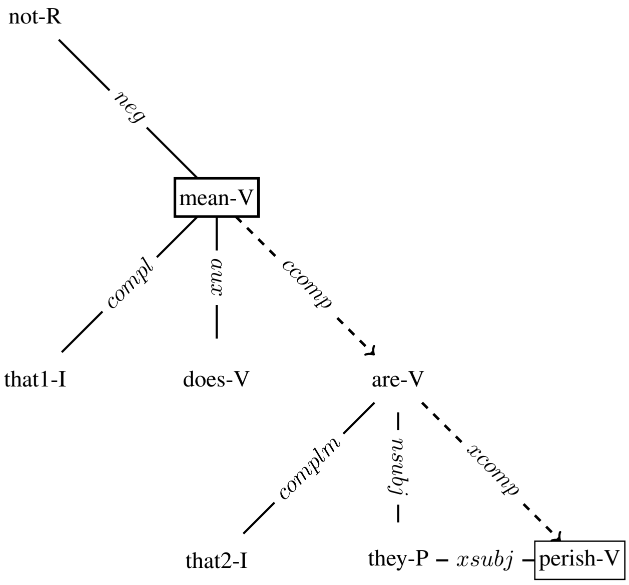
DEPENDENCY TREE TRAVERSAL	POS-GENERALIZED SUB-TREE	STANCE-ATTITUDE PROFILE
“‘No doubt about it—**the world has become more cynical and everything or so it seems.”**  seems-V nsubj advmod it-P doubt-N dep dep dep No-D cynical-J and everything-N nsubj aux cop world-N has-V become-V	*seems → N → cynical*	$S + A-$
“[...] but that does not mean that they are to **perish**.”  not-R neg mean-V compl aux ccomp that1-I does-V are-V complm nsubj xcomp that2-I they-P xsubj perish-V	*not-mean→V→ perish*	$S - A-$

Figure 5.4: POS-generalized sub-tree extraction from example (52a), which contains the extraposed stance marker *seems* and (53a), which contains a negated *for*-stanced word.

5.2.5 Stance and attitude lexical resources

In this section, we describe the stance and attitude lexicons used by our subtree extraction algorithm (`stanceWords` and `attitudeWords` in lines 11 and 15, respectively, of Figure 5.2). For our attitude lexicon, we used the MPQA subjectivity lexicon (Wilson and Wiebe, 2005). Since there does not exist a comparable resource for stancetaking lexis, we adopt the approach of Somasundaran and Wiebe (2010), who constructed a stance lexicon using spans of text from the MPQA corpus annotated for "positive (negative) arguing."

As described in Wilson and Wiebe (2005), the majority of the sentiment-bearing words included in the MPQA subjectivity lexicon were compiled in a semi-automated manner. A lexico-syntactic pattern learning algorithm was used to generate patterns learned from sentences classified as subjectivity-denoting. These extraction patterns were in turn used to identify more subjective sentences with their corresponding extraction patterns and the process is then iterated with the newly learned patterns. Performance was evaluated using a gold-standard set of sentences manually tagged for subjectivity and strength of subjectivity according to the annotation scheme described in Wilson and Wiebe. Words and phrases extracted using this approach were combined with other, publicly available sentiment lexicons (Stone et al., 1966; Hatzivassiloglou and McKeown, 1997) and tagged for *prior* or "out-of-context" sentiment polarity. A word bears a prior polarity if, absent any sort of context, it seems to evoke something positive or negative. The resulting lexicon contains 4911 positive, 2718 negative, and 430 neutral expressions. Only positive and negative expressions were used in the experiments reported here.

To construct a stance lexicon comparable to the attitude lexicon described above, we followed the approach of Somasundaran and Wiebe (2010), who describe a stance lexicon constructed using the manually annotated MPQA corpus. As described in Wilson and Wiebe (2005), and Wilson (2008), the MPQA corpus contains 15802 sentences extracted from 182 U.S. and foreign news services split across 10 topics. These topics include coverage of social and political issues that were divisive at the time of collection such as the Guantanamo Bay detention center and ratification of the Kyoto protocol. Annotators were asked to identify and tag text spans displaying one of a number of attitude types, such as positive and negative sentiment and positive and negative arguing. (54) provides examples of sentences containing (boldfaced) text spans annotated as, respectively, positive and negative arguing.

(54) a. Putin remarked that the events in Chechnya **"could be interpreted only in the context of the struggle against international terrorism."**

b. Officials in Panama **denied that Mr. Chavez or any of his family members had asked for asylum.**

The boldfaced text spans in (54) display the canonical features of stancetaking language: in both cases this language is used to argue *for* or *against* the likelihood of an accompanying proposition. In (54a), use of the restrictive degree adverb *only* has the effect of limiting the possibility reading of *could* to a single scenario in which the events in Chechnya are interpreted in the context of the struggle against international terrorism. *Could only* is here used to "[restrict] the truth value of the proposition either primarily or exclusively" to the part advocated by the speaker (Biber et al., 1999: 556). In (54b), *denied*, which falls under the rubric of *disclaim* verbs in Martin and White's (2005) taxonomy of evaluative lexis, casts doubt on the truth of the proposition contained in the clausal complement, *Mr. Chavez or any of his family members had asked for asylum*.

Reasoning that the strongest indicators of stance in an annotated text span occur in the first few words of that span, Somasundaran and Wiebe (2010) extracted the initial ngram (up to three) of each stemmed text span. Each extracted ngram is considered the arguing trigger expression while the remainder of the span is considered secondary. These expressions are stored and labeled as *candidate positive (negative) arguing trigger expressions*. In (54a), for example, the unigram *could*, the bigram *could be* and the trigram *could be interpreted* are all candidate positive-arguing trigger expressions. Candidate expressions will display considerable category overlap: *could* might appear as a positive-arguing expression candidate (since it occurs at the beginning of *could be interpreted*) but it might also appear as a negative-arguing candidate (since it could appear at the beginning of a text span tagged for negative arguing such as *could not be interpreted*). For this reason, each candidate receives a score indicating the likelihood that it is a positive arguing or negative arguing expression, calculated as the candidate's frequency of occurrence in a positive (negative) arguing span divided by its frequency of occurrence in the entire MPQA corpus, as given in (55) and (56).

(55)

$$P(\text{pro-arguing}|\text{candidate}) = \frac{\text{\#candidate is in a pro-arguing span}}{\text{\# candidate is in the MPQA corpus}}$$

(56)

$$P(\text{con-arguing}|\text{candidate}) = \frac{\text{\#candidate is in a con-arguing span}}{\text{\# candidate is in the MPQA corpus}}$$

The higher of these scores determines a candidate's final label as positive- or negative-arguing .

We replicated Somasundaran's approach with a Python script that used the annotation format scheme detailed in the MPQA documentation[1] to first collect all annotated spans in the MPQA corpus and to then locate those spans tagged *arguing-pos* or *arguing-neg*. This resulted in a list of 2378 spans tagged *arguing-pos* and 580 spans tagged *arguing-neg*. We then used the algorithm given in Somasundaran (2010, pg. 178) to extract the initial ngram (up to three) of each span. This resulted in a lexicon of 2166 *for*-stanced and 513 *against*-stanced ngrams.

After manually examining the resulting ngram lexicon, we found that its noisiest sections involved bigrams and trigrams. In the *for*-stance lexicon, for example, we find *in his*, *have re-elected*, and *the power of*, while in the *against*-stance lexicon we find *but we*, *realize that*, and *have been prompted*. We used only the unigrams in the lexicon since these appeared to be more

[1] http://www.mpqa.cs.pitt.edu/corpora/mpqa_corpus_2_0/mpqa_2_0_readme.txt

for-stance unigrams	*against*-stance unigrams
actually, assert, believe, claim, completely, conclude, consider, could, deem, demonstrate, evaluate, explain, insist, must, perhaps, probably, prove, regard, seems, show, signifies, stress, thought, typically, undeniably, will	allege, but, cannot, challenge, contravene, debate, deny, instead, lack, lie, nor, nothing, warn, without

Table 5.2: Examples of *for* and *against* unigrams from a stance lexicon constructed using the methodology of Somasundaran and Wiebe (2010). Additionally, metadiscourse markers from Hyland (2005) were added to the list of *for* unigrams.

reliably scored as *pro*-arguing or *con*-arguing. For example, *realize that* is included in the *against* lexicon, though most native speakers would likely consider this a *for* expression. However, by itself *realize* is correctly scored as *for*-stanced. We also pruned any unigrams that were obviously not stancetaking (*thanks, anybody, pre-election, suspicions*) resulting in an initial unigram lexicon of 336 *for* unigrams and 80 *against* unigrams.

To supplement this list, we used a selection of the metadiscourse markers listed in the appendix of Hyland (2005). Markers from the following categories were used: boosters (*clearly, decidedly*), hedges (*claim, estimate*), and engagement markers (*demonstrate, evaluate*). All of these markers were adjudged positively stanced by the criteria given in Martin and White (2005) and thus were added to the list of *for* unigrams. With Hyland's metadiscourse markers added to the initial lexicon, the final lexicon consists of 373 *for* and 80 *against* unigrams. Table 5.2 presents a selection from this final lexicon.

Now that all of the various steps and knowledge sources used to construct our POS generalized stance-attitude subtree feature have been described, we can diagrammatically summarize the steps used to construct this complex feature. Figure 5.5 summarizes the steps involved in creating the POS generalized stance-attitude subtree *can-V-true*.

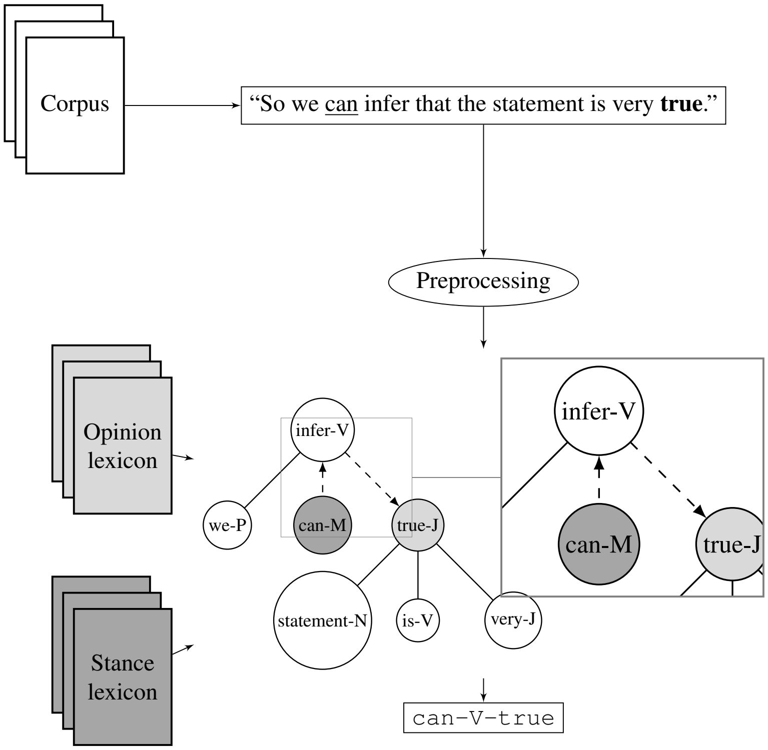


Figure 5.5: Scheme showing the creation of the POS generalized stance-attitude subtree *can-V-true* using dependency parsing, lexicon lookup, tree traversal, and POS generalization.

5.3 Prompt topic features

In this section, we describe the creation of a set of features motivated by related work in aspect-based Sentiment Analysis. We consider the task of linking stancetaking language in a given essay with aspects of one or more propositions contained in the prompt as analogous to the task in aspect-based Sentiment Analysis of linking targets of opinions in a given text, such as a product review, with aspects of the topic of that review. The feature representation scheme of Popescu and Etzioni (2005) consists of $\langle opinion, productaspect\rangle$ tuples, which capture the relationship between an opinion expression and some aspect of the global target of that opinion. However, we are interested in generating, for each essay, a set of essay words that bear a semantic relationship to words in the prompt.

As mentioned in section 5.1, there exist numerous approaches to capturing semantic relationships. The choice of one particular approach rather than another must be guided by the kind of semantic relationship we are interested in capturing between two given terms. Will hyponymic relations (*wickedness* is a hyponym of *evil*) of the kind provided by WordNet suffice? Or would a metric that captured distribution-based similarity provide us the semantic depth required to identify a relationship between terms such as *science* and *evolution* when used in an essay responding to the *Science* prompt?

Consider the *Money* prompt, given in (57) and a selection of sentences responding to this prompt given in (58).

(57) In the words of the old song,“Money is the root of all evil.”

(58) a. However, since the concepts dealt with here remain rather complex even if attempts are made to clarify and limit their scope, it is not within the range of this short essay to draw any definite conclusions on the subject.

b. You can’t be cured if you are ill, you can’t eat anything if you are hungry and you can’t sleep in a warm place.

c. No one in today’s world is able to exist without **money**.

d. **Rich** people can go any where they want to look for the cure of their diseases, whereas the **poor** don’t even be diagnosed as they can’t go to a doctor.

e. **Raskolnikov** killed the old woman because he decided that according to his theory such deed can be done.

Restricting our attention to the content words contained in the proposition, *money, root*, or *evil*, we can scan each of the sentences in (58) to discover potential semantic links between these content words and the words contained in a given sentence. If we do find such a link, we can be confident that the sentence contains subject matter that deals directly with the proposition, *Money is the root of all evil*. In (58a,b) we do not find any obvious semantic links between the words in each of these sentences and the words contained in the proposition. We can assume, then, that neither of these sentences deals directly with the proposition in the prompt. In (58c), on the other hand, we find the word *money* which directly links this sentence with the prompt since *money* is also found in the proposition contained in the prompt.

In (58d) and (58e) we find more complex examples of semantic links between words in the sentence and content words in the prompt. Intuitively, in (58d), *rich* and *poor* deal with the subject of money since both these words are definitionally related to *money*. To have a lot of money is to be rich; to have very little money is to be poor. More generally, we can say that *rich* and *poor* share the same "semantic field" (Wierzbicka, 1996) as *money*.[2] In (58e), we find the name *Raskolnikov* which, unlike *rich* and *poor*, does not bear the sort of lexical relationship to *money* or *evil* that can be established by consulting thesauri such as WordNet. Instead, *Raskolnikov* is part of the same semantic field as *evil* and *money* by virtue of the subject matter of the novel *Crime and Punishment*, which deals with the murder of a pawnbroker for money and relativistic notions of evil. Both (58d) and (58e), then, contain words in the same semantic field as the language of the proposition, and thus likely deal with aspects of that proposition.

5.3.1 The Wikipedia Link-based Measure (Witten and Milne, 2008)

While knowledge-based resources such as WordNet and Lin's proximity thesaurus might be able to capture lexical groupings such as *money* and *rich*, cultural kinds such as *Raskolnikov* are not part of the same lexical family as any of the words in (57). To capture relationships between words grouped within the same semantic field by cultural association, we must make use of a corpus-based metric of word similarity. This metric would capture the relationship between lexically unrelated words as a function of their distributional similarity or, in the case of web- or Wikipedia-based metrics, hyperlink structure. Ideally, the corpus upon which this metric is based should deal with a vast number of topics since cultural kinds are grouped by association rather than by any principled measure of semantic similarity (Wierzbicka, 1996, pg. 172). Web-based

[2] Our understanding of semantic fields is drawn from Wierzbicka (1996) who distinguished between semantic fields dealing with "natural kinds" which have discrete taxonomies (*oak, willow, birch*, and *palm* are all subsumed under the supercategory of *tree*) and semantic fields dealing with "cultural kinds" (Wierzbicka, 1996, pg. 172). Wierzbicka argues that, unlike natural kinds, which are members of a hierarchically structured, discrete taxonomy, category membership for fields containing cultural kinds, such as *cup*, *mug*, and *jar*, is a more or less arbitrary affair, since speakers can generate chains of associations linking disparate cultural kinds that do not fall under a discrete superordinate category. Thus, while some speakers might place *cup, mug, bottle, jar,* and *bucket* under the category of "containers," some speakers might also feel that *bucket* and *tub* are related, though *tub* is not generally thought of as a kind of container.

semantic similarity metrics such as Cilibrasi and Vitanyi's (2007) Normalized Google Distance (NGD) exploit the vast amount of text on the internet and Google hit counts to associate related words across a large number of topics. Unfortunately, large scale ($\geq$ 100 word pairs) use of NGD is prohibitively expensive for researchers due to Google's search API fees.

Witten and Milne (2008) used the inter-link structure of articles in the English version of the collaboratively edited online encyclopedia Wikipedia[3] to create an alternative to NGD, the Wikipedia Link-based Measure (WLM). The typical Wikipedia page contains a large network of cross-references in the form of internal (connected to another Wikipedia page) and external (connected to a page outside of Wikipedia) hyperlinks. Wikipedia's documentation instructs page editors to include links to related topics, so that an article on *planes* should link to *fixed wing aircraft, airplane*, and *aeroplane*. In this way, Wikipedia's internal link structure can be viewed as a kind of naturally-occurring ontology. Milne and Witten experiment with two forms of WLM, the first measuring the relatedness of two terms as a function of outgoing links (links to other Wikipedia pages) and the other measuring relatedness as a function of incoming links, or "backlinks" (links from other Wikipedia pages). We use the latter measure since, as Milne and Witten report, this measure beat the former by six percentage points in experiments using three ground-truth semantic similarity datasets provided by Rubenstein and Goodenough (1965), Miller and Charles (1991), and Finkelstein et al. (2001).

Given the *Money* prompt, with its associated content words, *money, root*, and *evil*, suppose we are given a clause from an essay response that contains the content word *capitalism* and another clause that contains the content word *Raskolnikov*. Using just the prompt word *money* in this example, we could determine whether *capitalism* or *Raskolnikov* has a closer semantic relationship to *money* by calculating the WLM of *money/capitalism* and *money/Raskolnikov.* Milne and Witten define the WLM as

[3] http://en.wikipedia.org/wiki/Main_Page

$$wlm(a,b) = \frac{log_{10}(max(|A|,|B|) - log_{10}(|A \cap B|)}{log_{10}(|W|) - log_{10}(min(|A|,|B|))}$$

where *a* and *b* are Wikipedia article titles (e.g., the articles for *money* and *capitalism*), *A* and *B* are the sets of articles that backlink to *a* and *b*, and *W* is the count of all articles currently contained in Wikipedia (as of this writing, $\sim$ 4.3 million). As given, if $wlm(a,b) = 0$ then *a* and *b* are as semantically similar as possible and if $wlm(a,b) \geq 1$ they are semantically dissimilar. For ease of interpretation, we subtract all WLM scores from 1, so that that a score of 1 means that *a* and b are as similar as possible. In our implementation of the measure, if there is no intersection between article backlinks (i.e., $(|A \cap B|) = 0$), then that particular *a*,*b* pair is not evaluated.

The WLM scores for *money/capitalism* and *money/Raskolnikov* are .59 and .34, respectively. This fits with the general intuition that *capitalism* is more naturally paired with *money* than is *Raskolnikov*. [4]

The WLM was used to score essay words relative to content words contained in the prompt to which the essay is responding. The essay words scored were content words contained in a proposition targeted by a stance word, where propositions were identified using the criteria described in section 3.6.2. Essay words that received a WLM score ≥ 0 relative to a prompt word are considered aspects of the proposition. The resulting feature set consists of an unordered collection of stemmed topic words.

[4] All WLM scores were calculated using a Python script that makes calls to Wikipedia's API, retrieves the backlinks associated with each term, and then calculates the WLM scores using the formula described above. For the current example, as of 9/28/13, the number of backlinks to *capitalism* is 6895, the number of backlinks to *money* is 2390, the number of backlinks to *Raskolnikov* is 33, and the total number of Wikipedia articles is 4,337,306. The number of backlinks common to *money/capitalism* and *money/Raskolnikov* is 302 and 1, respectively. The WLM score for the first of these pairs, then, is

$$wlm(\text{money, capitalism}) = \frac{log_{10}(6,895) - log_{10}(302)}{log_{10}(4,337,306) - log_{10}(2,390)} = .41$$

Subtracted from 1, this is .59. The WLM for *money/Raskolnikov* is

$$wlm(\text{money, Raskolnikov}) = \frac{log_{10}(2,390) - log_{10}(1)}{log_{10}(4,337,306) - log_{10}(33)} = .66$$

Subtracted from 1, this is .34

5.3.2 Scoring topic words using the Wikipedia Link-based Measure

For each essay in the subset of the corpus, a set of stemmed topic words with WLM scores ≥ 0 was created using the following procedure:

- **Step 1**. Using the stance lexicon described in section 5.2.5, all stance words in a given essay were identified.

- **Step 2**. The `clauseRest` function used by the sub-tree extraction algorithm (Figure 5.2) was used to identify propositions that meet the clausal restrictions described in section 3.6.1.

- **Step 3**. For each content word in the prompt to which the essay is responding, a WLM score was calculated relative to all content words contained in the proposition identified in Step 2.

- **Step 4**. Many essay words received WLM scores ≥ 0 for more than one prompt word. In such cases, the highest WLM score is assigned to that essay word.

- **Step 5**. All essay words are stemmed.

Step 4 requires further explanation. It will often happen that an essay word will receive a WLM score ≥ 0 relative to more than one prompt word. For example, the essay word *law*, when it appears in an essay responding to the *Science* prompt, has the WLM scores given in Table 5.3. In this case, the word *law* would receive the highest WLM score from this list, .8175.

We also found many essay words with duplicate WLM scores. These duplicate scores are the result of our use of the Wikipedia API's *redirect* function which redirects all morphological variants of a given word to a single article. Thus, when making a call to the Wikipedia API for backlink counts for words such as *morality, moralize*, and *moral*, the redirect function will redirect such calls to the single page *morality* and so each of these words will have the same number of backlinks associated with *morality*.

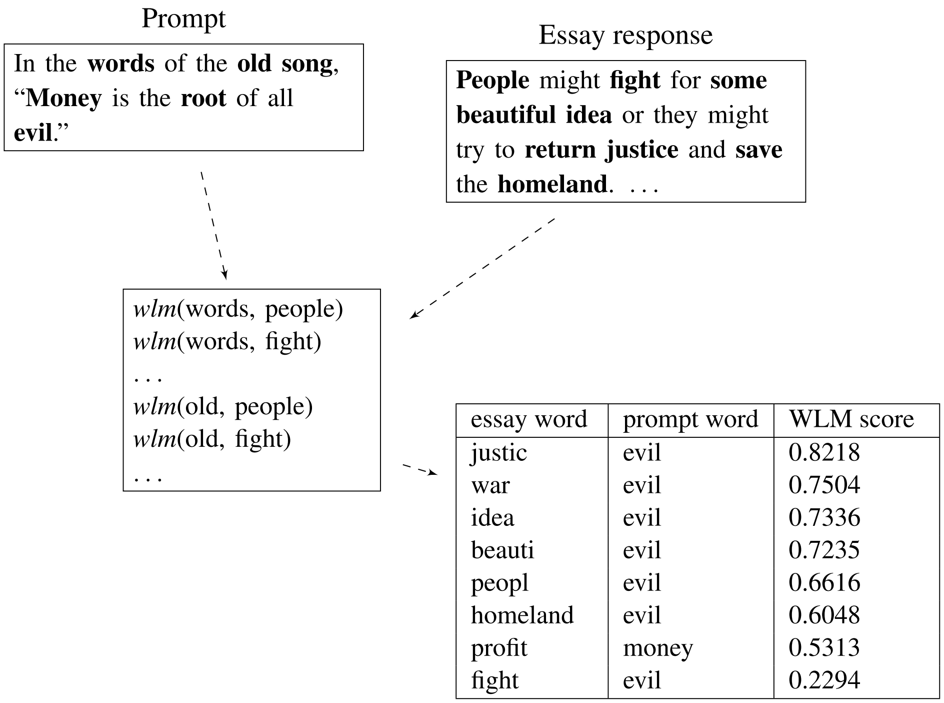
All 5 steps can be illustrated using a segment of text responding to the *Money* prompt, given in (59).

ESSAY WORD	PROMPT WORD	WLM SCORE
law	science	0.8175
	technology	0.7619
	dominated	0.6399
	imagination	0.6377
	world	0.6199
	people	0.6022
	industrialization	0.0
	modern	0.0
	dreaming	0.0

Table 5.3: WLM-scored prompt word/essay word pairs for the word *law* as it appears in an essay responding to the *Science* prompt.

(59) People might fight for some beautiful idea or they might try to return justice and save the homeland. Certainly, the beautiful idea is nothing but a cover-up. We don't have to go far for the examples. There is an excellent example before our very eyes. The Chechen war. The war burst out because somebody needed to laundry a huge sum of money. The war lasted because it was profitable to make money of it. The war ended (they say we even won) because it became profitable to make money off peace.

Examining this text segment, it is obvious that only a subset of these sentences deals directly with some aspect of the proposition contained in the *Money* prompt. Other sentences are fragments of a larger argument (*We don't have to go far for the examples*) or have been truncated for rhetorical effect (*There is an excellent example before our very eyes. The Chechen war*). After applying the procedures described in steps 1 and 2, all content words contained in the propositions targeted by identified stance words in (59) are identified. For each of these content words, a WLM score is calculated relative to the content words contained in the *Money* prompt (step 3). If an essay word received a WLM score for more than one prompt word, the highest WLM score was retained (Step 4). After stemming the resulting list of words (Step 5) using the Snowball stemmer of Porter (2001), a final list of prompt topic words for this text segment was generated. These steps are summarized in Figure 5.6. In Table 5.4, we provide the top ten, rank-ordered WLM-scored prompt word/essay words for each of the seven prompts.



In the **words** of the **old song**, “**Money** is the **root** of all **evil**.”

People might **fight** for **some beautiful idea** or they might try to **return justice** and **save** the **homeland**. . . .

wlm(words, people)
wlm(words, fight)
. . .
wlm(old, people)
wlm(old, fight)
. . .

essay word	prompt word	WLM score
justic	evil	0.8218
war	evil	0.7504
idea	evil	0.7336
beauti	evil	0.7235
peopl	evil	0.6616
homeland	evil	0.6048
profit	money	0.5313
fight	evil	0.2294

Figure 5.6: Creation of a set of stemmed, WLM-scored topic words for the text segment given in (59).

Armies prompt

essay word	prompt word	WLM score
polic	military	0.8639
bureaucraci	military	0.8475
constitut	military	0.8467
contract	military	0.8443
product	service	0.8440
lawyer	military	0.8337
law	military	0.8175
naval	armies	0.8141
navi	armies	0.814
sergeant	soldiers	0.8082

Degrees prompt

essay word	prompt word	WLM score
colleg	university	0.8981
undergradu	university	0.8484
postgradu	university	0.8409
condit	value	0.8234
statement	value	0.8179
direct	degrees	0.8086
fact	theoretical	0.8044
object	value	0.8032
chang	value	0.8027
observ	theoretical	0.8005

Feminism prompt

essay word	prompt word	WLM score
girl	women	0.8785
feminin	women	0.8718
womanli	women	0.8718
antifemin	women	0.8652
patriarchi	women	0.8499
patriarch	women	0.8499
man	women	0.8197
men	women	0.8197
leav	harm	0.8083
hurt	harm	0.8083

Marx prompt

essay word	prompt word	WLM score
nietzsch	marx	0.8909
radio	television	0.8914
hegel	marx	0.8767
engel	marx	0.8755
socialist	marx	0.8711
judaism	religion	0.8698
marxism	marx	0.8625
marxist	marx	0.8625
god	religion	0.8597
communism	marx	0.8568

Money prompt

essay word	prompt word	WLM score
virtu	evil	0.8443
happi	evil	0.8417
suffer	evil	0.8391
decent	evil	0.8365
immor	evil	0.8365
moral	evil	0.8365
indec	evil	0.8365
conscienc	evil	0.8356
principl	evil	0.8346
currenc	money	0.8337

Prisons prompt

essay word	prompt word	WLM score
justic	society	0.8346
rape	society	0.8337
rapist	society	0.8337
incest	society	0.8328
prostitut	society	0.8309
obscen	society	0.8299
cultur	society	0.8260
revolut	society	0.8218
revolt	society	0.8218
anarch	society	0.8208

Science prompt

essay word	prompt word	WLM score
tool	technology	0.8632
philosophi	science	0.8523
logic	science	0.8491
psycholog	science	0.8426
know	technology	0.8391
knowledg	technology	0.8391
known	technology	0.8391
construct	technology	0.8390
machin	technology	0.8383
nourish	technology	0.8374

Table 5.4:   Top ten, rank-ordered WLM scores for all seven prompts.

Chapter 6

Essay-level stance classification experiments

6.1 Related work

The work most directly related to the document-level stance classification experiments described in this chapter is Somasundaran and Wiebe (2010), Anand et al. (2011), Walker et al. (2012a), and Hasan and Ng (2013a,b). These researchers present various document-level stance classification models trained and tested on data scraped from online debate forums. Working with a set of $\sim$2200 debate posts dealing with topics such as gun rights, healthcare reform, and the existence of God, Somasundaran and Wiebe (2010) present a series of classification experiments using lexicon-based features. A significant aspect of their system is its recognition that stancetaking lexis exists as a separate category of evaluative language, quite apart from the sentiment-bearing lexis that has been the focus of most Sentiment Analysis work. A stance lexicon was constructed using ngrams extracted from text spans annotated for positive (negative) arguing in the MPQA corpus.

Additionally, Somasundaran and Wiebe incorporated information regarding the targets of both opinion and argument stance into their feature sets. Three different feature sets were developed for their classification experiments. The first set is designed to capture stancetaking language and its targets, while the second set is designed to capture opinionated langauge and its targets. The third feature set combines the first two sets. Somasundaran and Wiebe's approach to incorporating information about the targets of stancetaking language involved first identifying the dominant stance of a given sentence. This is determined by summing the likelihood scores associated with any stance words found in the sentence and considering the higher-scoring sum as the dominant stance of that sentence. Likelihood scores are calculated as the frequency of that stance word's occurrence in the positive (negative) arguing spans identified by MPQA annotators, divided by its frequency of occurrence in the entire MPQA corpus (in the current study, these formulae are given as (55) and (56) in section 5.2). This serves as a rough measure of the word's reliability as an indicator of positive (negative) arguing.

The label associated with each sentence's dominant stance was then appended to every content word in the sentence—for example, *abortion should be banned* becomes `<abortion-an should-an be-an banned-an>`, where *an* = *arguing negative*. For the creation of the second, opinion features set, the dominant opinion of a sentence was determined using the Vote and Flip algorithm of Choi and Cardie (2009) (cf. section 8.2.2) in combination with the MPQA subjectivity lexicon. A similar word labeling procedure was used to append positive (negative)-sentiment tags to all content words in the sentence. The resulting text is thereafter transformed into a frequency-valued bag-of-words feature set and various combinations of these two feature sets were used as training data for an SVM learner. The resulting system achieves a somewhat low accuracy of 63% across all debate topics. This score is arguably the result of an overly coarse-grained approach taken to identifying stance targets (cf. section 5.2.1).

The system of Anand et al. (2011) was trained on a debates corpus of 4772 posts. Reasoning that the head-modifier relations returned by the Stanford dependency parser could serve as a coarse-grained method of capturing stance target information, Anand et al. (2011) first reduce each debate post to *(opinion, target)* dependency tuples. To increase feature redundancy, these tuples were then generalized to tuple members' POS. An accuracy of .64 was achieved using ngrams, generalized dependencies, and various other features including sentence length and pronominal form occurrence. More recently, Hasan and Ng (2013a,b) experimented with the feature set of Anand et al.,"extra-linguistic constraints" such as the stance of an immediately preceding post (an approach to capturing inter-post relations is also presented in Walker et al., 2012), the writer's stance regarding other debate topics, and features based on frame semantics. The highest accuracy reported was .74 for a set of debates on abortion. Domain-general results were not reported.

An issue left unaddressed by these researchers is whether online debate data are truly representative of argumentative text. The language of these debates is often highly emotion-laden (a feature of opinion-bearing, rather than stancetaking language), often consists of only one or two sentences, and displays little of the organizational features typical of argumentative text. In this chapter, we address this point using quintessentially argumentative data—student argumentative essays. Additionally, the approach to capturing stance target information described in Somasundaran and Wiebe, Anand et al., and Hasan and Ng is coarse-grained, resulting in noisy examples of stance targets. By contrast, our approach to stance target identification, as described in section 5.2.4, uses a lexicon-based method to find stance words and their targets, and restricts the identified stance targets to targets that meet certain syntactic criteria. The result is a smaller, but more fine-grained feature set than that of Somasundaran and Wiebe, Anand et al., and Hasan and Ng.

6.2 Learning algorithms

Our classification experiments assume a supervised classification framework, which consists of a document space and a set of classes, where the document space is some high-dimensional space of features, such as word counts, and the classes are the document labels we are interesting in predicting (in our case, the labels *for* or *against*). Given a training set of hand-labeled documents, such as our set of annotated ICLE essays, the goal of a learning algorithm in a supervised classification framework is to learn some classification function that maps these documents to classes.

In this section, we describe the two learning algorithms chosen for the classification experiments reported here, Naive Bayes (NB) and support vector machines (SVM).

Our choice of learning algorithms was motivated by, (a) the relatively small size and restricted domain of our corpus, (b) the large number of features used in our experiments and, (c) our interest in evaluating the validity of the linguistic generalizations motivating our features. For text classification tasks involving small, single-domain corpora, it is recommended practice (Manning et al., 2008, pg. 336) to make use of so-called *high bias* algorithms such as NB. The term *bias* is part of a concept that is central to machine learning, the *bias-variance tradeoff* (Geurts, 2005, pg. 749). Given some document d and a class c, *bias* is the difference between the true conditional probability of d being in class c and the predictions made by a classifier averaged over multiple training sets. Of course, we have only a single training set, our selection of ICLE essays. However, suppose we were to create multiple models over different versions of the training set. Given the randomness inherent in these different sets, there will be some range of predictions made by the classifier. If the classifier consistently misclassified d across these different sets, we can describe the classifier as having high bias. On the other hand, if different training sets produce small errors in the classifier's predictions for d or if the classifier makes consistently correct predictions for d across training sets, the classifier can be described as displaying low bias

Variance can be viewed as measure of how inconsistent a classifier is, averaged over multiple training sets. Imagine, again, that multiple models are created across different training sets. A classifier can be described as having high variance if its predictions for d vary greatly across these different sets. The tradeoff between bias and variance can be explained in terms of model *over(under)fitting*. An overfit model, which fits a given training set perfectly, has low bias (since its predictions are not consistently wrong) but high variance (since its predictions can vary greatly with new training data). An underfit model has high bias since it stubbornly persists in its hypothesis regarding d across different training sets and also low variance since new training data has little effect on the (correct or incorrect) classification decisions made by the classifier.

NB is an inherently high bias/low variance learning algorithm (Geurts, 2005, pg. 753). It is high bias since its hypotheses are limited to a handful of conditional probability estimates which are not affected by small changes in the training data. At the same time, it is low variance precisely because of its use of a small set of inflexible parameters. For this reason, NB is amenable to text classification situations involving a relatively small amount of single domain data (Manning et al., 2008, pg. 336) while more flexible, low-bias learning algorithms such as K-nearest neighbors (kNN) are better suited to situations involving large amounts of mixed-domain corpora. Since we are not concerned with generalizing this model beyond the domain of our relatively small set of ICLE essays, we use NB to generate a high-accuracy model which is not necessarily generalizable to out-of-domain text.

Additionally, given our general interest in evaluating the relative performance of the linguistically motivated features described in chapter 5, our choice of NB is motivated by the fact that the classification model generated by this learning algorithm can be easily interpreted. While several studies comparing different learning algorithms consistently show kNN outperforming NB across various text-classification tasks (Yang, 1999; Yang and Liu, 1999), kNN models are generally difficult to interpret and so such models would provide us with few insights into the validity of the linguistic generalizations motivating our feature set.

Our second learning algorithm, the SVM, is easily interpreted and achieves high-accuracy across a range of text classification tasks, as reported in both early and recent comparative studies (Yang and Liu, 1999; Caruana and Niculescu-Mizil, 2006). Additionally, SVMs are ideally suited to very high-dimensional features spaces such as our classification features sets (our feature set combining stance-attitude subtrees with all topic words with WLM scores ≥ 0, for example, has 1225 unique attributes).

We first describe the NB learning algorithm and the two models of NB utilized in the experiments reported here, the *multinomial* model of NB (NBM) and the *multivariate Bernouli* model of NB (NBMB). The difference between these two classification models turns on the feature estimates used when making its calculations. The multinomial model calculates its estimates by counting the number of times a feature occurs in a document while the multivariate Bernouli model calculates its estimates by tracking binary occurrence/non-occurrence information for these features. McCallum et al. (1998) and Manning et al. (2008) present a comparative analysis of these two models for text classification tasks. Much of the notation used in this section is adapted from this work.

NBM first calculates the probability that some document d is in class c by multiplying the *prior probability* $P(c)$ of that document occurring in class c by the conditional probability of some term t_k occurring in a document labeled class c. This formula is given as (60).

(60)

$$P(c|d) = P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

In (60) the parameter $P(c)$ can be estimated as the number of documents in class c over the total number of documents N, as in (61).

(61)

$$P(c) = \frac{N_c}{N}$$

The conditional probability parameter $P(t_k|c)$ can be estimated as the frequency of occurrence of some term t in all documents of class c, as in (62).

(62)

$$P(t_k|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

In (62), T_{ct} is the total number of occurrences of term t in all training documents of class c and the denominator $\sum_{t' \in V} T_{ct'}$ is the sum of all occurrences of any given term t' in vocabulary V for all training documents of class c—the denominator, in other words, is the total vocabulary length for training documents of class c.

Making predictions using the formula in (60) involves picking the most likely, or *maximum a posteriori* (MAP) class c given a test document d. This step can be represented as the formula c_{map} in (63).

(63)

$$c_{map} = \arg\max_{c \in \mathbb{C}} P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

In general, every text collection will generate a certain number of 0 values for some conditional probability $P(t_k|c)$ for some term t. For example, the word *disagree* might occur only in the *against* class of training documents in the ICLE corpus. A test document of class *for* containing the word *disagree* would then receive a conditional probability of 0 since $P(disagree|\textit{for}) = 0$ is being multiplied by each of the conditional probabilities of the other terms found in the test document. To correct for this "data sparseness" problem, most implementations of NB employ *Laplace* or *add-one smoothing* which adds one to each of the counts in (62). With add-one smoothing incorporated, (62) can be rewritten as (64).

(64)

$$P(t_k|c) = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + |V|}$$

In (64), $|V|$ is the count of all terms in the vocabulary V.

Under the multinomial model, the conditional probability $P(t_k|c)$ is computed by counting the number of occurrences of t across all document of class c. By contrast, under the Bernouli model, multiple term occurrence is not captured and $P(t_k|c)$ is calculated by counting the number of documents of class c containing term t. The difference between these two models, then, turns on the assumptions each model makes regarding the distribution of $P(t_k|c)$. The multinomial model assumes that each document d is drawn from a multinomial distribution with the number of independent trials equal to the length of d. The Bernouli model, on the other hand, assumes that d is distributed according to a multivariate Bernouli distribution. Thus, d can be viewed as a set of independent Bernouli trials, with success or failure depending on whether the word does or does not occur in the document.

Importantly, both NBM and NBMB make a *conditional independence assumption* when estimating their respective parameters. In fact, this is the signature assumption of the NB framework in the context of text classification tasks: each term t is considered positionally independent of any other term. While it is clear that conditional independence is empirically false for natural language data (*the* is far more likely to precede a noun such as *man* than to follow such a noun), this assumption greatly reduces the number of parameters that must be estimated since the number of conditional probabilities $P(t|c)$ that must be calculated will always be $|V| \times |C|$ where $|C|$ is the number of classes in the training data. By contrast, without the positional independence assumption, the number of conditional probabilities to be estimated will be exponentially larger since each possible combination of terms in V relative to each class in C must be calculated.

Our second classification algorithm is the SVM learning algorithm. In the present context, it is not necessary for the reader to fully understand the mathematical details underlying SVMs. However, a conceptual grasp of SVMs along with an understanding of the math underlying the classification function generated by this algorithm will be beneficial when reading through our evaluation of experiment results in section 6.4. To that end, we describe the following facets of SVMs: *large-margin classification*, *support vectors*, *kernels*, and the resulting *classification function* learned by an SVM model. We provide a conceptual description of the first three items in this list and will discuss a portion of the math underlying the last item. Readers interested in accessible introductions to SVMs, with no derivations of the underlying mathematical concepts, are referred to Burges (1998) and Manning et al. (2008).

Intuitively, given a two-class classification problem such as our stance classification task, if the two classes are linearly separable in an n-dimensional feature space (in our case, n equals the number of terms in our feature sets), one should pick a boundary separating the two classes that maximally separates them. The larger the space between the classes, the less likely it is that the classifier will err in its predictions regarding data points (in our case, essays) close to the boundary on either side. This is the intuition underlying *large-margin* classifiers such as SVMs. Further, only a subset of data points will play a role in determining the position of the boundary, or *decision plane*—these are the *support vectors*. Figure 6.1 illustrates these concepts. In Figure 6.1 the five support vectors that serve to determine the position of the boundary between classes are aligned along the margins of the decision plane.

Of course, real-world data is rarely linearly separable. How then can we transform the data in such a way that it is linearly separable? *Kernel* functions serve to re-cast points in the training data into a linearly separable format by mapping these points to a higher-dimensional feature space. In the experiments reported here, we used a Radial Basis Function (RBF) kernel. This kernel performed slightly better than an alternative, linear kernel. At this point, the SVM algorithm uses an optimization procedure to find the maximum-margin hyperplane in this transformed feature space. Weka's implementation of the SVM algorithm uses the Sequential Optimization Algorithm of Platt et al. (1998) for this step. After optimization, the learned classification function uses only the identified support vectors to make its decisions—all other training data is discarded. Given an essay represented as a vector of test data points $\vec{x}$ the classification function $f(\vec{x})$ projects this point onto the region of the space perpendicular to the maximum margin hyperplane (i.e., the hyperplane normal). The sign of the resulting function determines which class to assign to this test point. The classification function $f(\vec{x})$ is defined in (65).

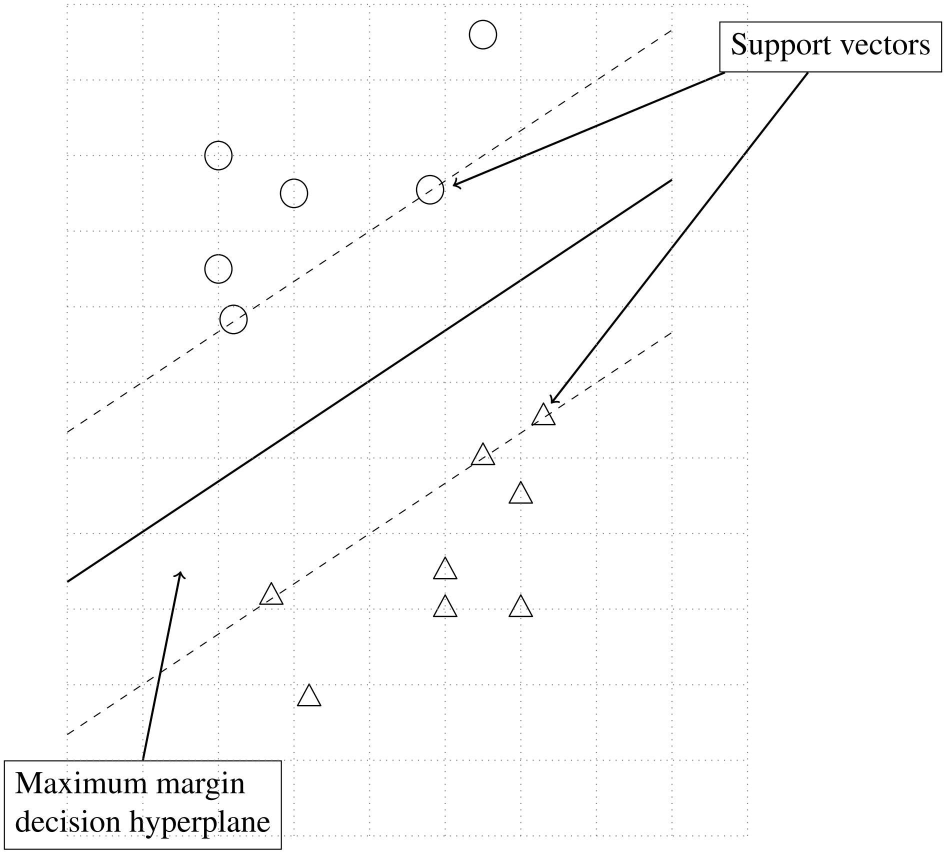


Figure 6.1: An illustration of the large-margin framework implemented by SVMs. The decision hyperplane is determined by the five support vectors aligned along the border of the margin. The margin is optimized in such a way that the two classes of data are maximally separated.

(65)

$$sign(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b)$$

(65) takes as input points i through m in the training data. The coefficient α is the value associated with the support vectors learned from the optimization step described above and y_i is a class label for each point, which can have a value of -1 or 1 depending on the class. As mentioned, the kernel function K in our case is an RBF kernel. As input, K takes an n-dimensional vector x_i and the n-dimensional vector x that we are interested in classifying. Finally, b is a scalar value, again learned in the optimization step.

6.3 Experiment setup

All experiments were conducted using using the Weka machine learning toolkit (Hall et al., 2009). 1135 essays were separated from the set of 1320 described in section 4.2. The 1135 essays served as the training/test set for the experiments reported here, with the remainder of the 1320 essays set aside for use as a development set. Of the 1135 essays, 498 were tagged *for* by annotators and 637 were tagged *against*, giving us a majority-class baseline of 56%. In order to make the task more challenging, we supplemented this rather low baseline with two alternative model baselines, a bag-of-words model and a model based on the best-performing set of features described in Somasundaran and Wiebe (2010). In the context of document-level text classification tasks, bag-of-words models represent documents as a vector of word frequencies. As Pang and Lee (2008) observe, such simplistic models are surprisingly hard to beat for Sentiment Analysis tasks and thus can serve as a high baseline for classification models based on more sophisticated feature sets.

Our second baseline is a model constructed using the highest-accuracy feature set described in Somasundaran and Wiebe. This feature set consists of a combination of stance polarity-tagged stemmed content words and opinion polarity-tagged stemmed content words. While the feature sets of Anand et al. (2011) and Hasan and Ng (2013a,b) are somewhat more sophisticated than that of Somasundaran and Wiebe, we used the latter-most feature set as a baseline since it incorporates information regarding stance targets and makes use of a larger version of the stance lexicon also used here. By comparing our two systems, we can get a sense of whether our approach to incorporating stance target information represents an improvement over that of Somasundaran and Wiebe.

In section 5.2.1, we described how Somasundaran and Wiebe constructed a feature set of stance polarity-stemmed content words. Constructing a feature set consisting of stemmed opinion polarity-tagged content words follows a similar procedure. For each sentence in a particular debate post, any opinion-bearing or neutral words are identified using the MPQA subjectivity lexicon (Wilson and Wiebe, 2005). Rather than using majority counts of opinion-bearing words to determine the polarity to assign to a particular sentence, the Vote and Flip algorithm of Choi and Cardie (2009) is used (cf. section 8.2.2 for a description of this algorithm). The Vote and Flip algorithm is designed to account for single or multiple negators in a given expression containing opinion-bearing language (e.g., *not a great movie, but not the worst either*), and to flip or not flip the expression's dominant polarity accordingly, or to return a dominant polarity of *neutral*. Somasundaran and Wiebe extend the application of this algorithm to determine the dominant opinion polarity of whole sentences rather than just expressions. Once the dominant polarity of each sentence is determined, tags of "+" (=positive opinion), "-", (=negative opinion) or "=" (=neutral) are appended to each stemmed content word in that sentence. We replicated this procedure for each of our 1135 ICLE essays. A typical sentence from an ICLE essay tagged in this manner is given in (66).

(66) <there, are-,quit-, a, few-, peopl-, who, think-, that, higher-, educ-, is-, unnecessari-, and, that, there, is-, no, point-, in, studi-, at, univers->.

These opinion polarity-tagged features were then combined with the stance polarity-tagged features described in section 5.2.1 to create our second baseline model.

Given the somewhat small size of the dataset we elected to use 10-fold cross-validation for all training and testing. We experimented with three different feature representations. The first feature representation uses the stance-attitude feature representation framework described in section 5.2.4 to represent each essay as an unordered set of POS-generalized stance-attitude subtrees. A segment of one such essay representation (responding to the Prisons prompt) is given in (67).

(67) `<not-certain-N-outdat, not-should-V-outdat, should-accept, not -will-difficult, not-although-difficult>`

The second feature representation represents each essay as an unordered collection of stemmed, WLM-scored topic words extracted using the procedure described in section 5.3.2. A sample of a set derived from an essay responding to the *Armies* prompt is given in (68).

(68) `<shot, kill, nato, nation, govern, militari, soldier, war, tank, servic, general, enlist, rifl>`

The third feature representation combines the first two feature sets.

For experiments involving NBM and SVMs, all feature counts for both training and test data were normalized using Weka's data normalization function. This function works in the following manner. Given a set of documents D, the length of each document $d \in D$ is the vector length of d. This is calculated in (69).

(69)

$$vl = \sqrt{t_1^2 + t_2^2 \ldots + t_n^2}$$

In (69), t is some term in our feature set. Each term t in d is normalized by first calculating the average of all vl's in the training and test sets, which we denote as $\overline{vl}$. Each term in a document is then normalized by dividing $\overline{vl}$ by vl and multiplying this value by the number of occurrences of t. The normalized value for terms that occur multiple times in a single document is calculated by summing the normalized values associated with that term in that document The formula for the normalized value of a given term t, then, is (70).

(70)

$$t_{normalized} = n\left(\frac{\overline{vl}}{vl}\right)$$

In (70), n is the number of times the term occurs in d.[1]

6.4 Experiment results

We measure the performance of the NBM, NBMB and SVM learning algorithms across all three feature representations using metrics that have become standard in the literature: *Accuracy*, *Recall*, *Precision*, and *F-measure*. Accuracy, given in (71), is simply the percentage of correctly classified instances generated by the model.

[1] To illustrate, suppose we are given two documents reduced to feature sets:

essay 1	`<should, not-punish>`
essay 2	`<should, might-V-benefit, should>`

We use the formula in (69) to calculate the vector lengths of essays 1 and 2:

essay 1	`<should, not-punish>` = $\sqrt{1^2+1^2} = 1.414213562$
essay 2	`<should, might-V-benefit, should>` = $\sqrt{2^2+1^2} = 2.236067977$

The mean length $\overline{vl}$ is $\frac{1.414213562+2.236067977}{2} = 1.82514$. We use these scores to calculate the following normalized values for all terms in essays 1 and 2 using formula (78):

essay 1	`should` = $(1)(\frac{1.82514}{1.414213562}) = 1.290569414$, `not-punish` = $(1)(\frac{1.82514}{1.414213562}) = 1.290569414$
essay 2	`should` = $(2)(\frac{1.82514}{2.236067977}) = 1.6324555316$, `might-V-benefit` = $(1)(\frac{1.82514}{2.236067977}) = 0.816227765$

If the terms have been properly normalized, each document should have a vector length equal to $\overline{vl}$:

essay 1	$\sqrt{1.290569414^2 + 1.290569414^2} = 1.82514$ ✓
essay 2	$\sqrt{1.6324555316^2 + 0.816227765^2} = 1.82514$ ✓

(71)

$$Accuracy = \frac{\textit{true positives}}{\textit{true positives} + \textit{false positives} + \textit{true negatives} + \textit{false negatives}}$$

In a classification setting, Precision, given as (72), is a measure of the classifier's *exactness*—of the predictions made by the classifier, what percentage were correct? A higher precision means fewer false positives were returned by the classifier since the classifier has predicted correctly much of the time.

(72)

$$Precision = \frac{\textit{true positives}}{\textit{true positives} + \textit{false positives}}$$

Recall, given as (73), is a measure of the classifier's *sensitivity*—of the predictions made by the classifier, what percentage were supposed to be correct? In the case of Recall, a higher score means fewer false negatives since the predictions that were supposed to be correct were in fact correct.

(73)

$$Recall = \frac{\textit{true positives}}{\textit{true positives} + \textit{false negatives}}$$

The trade-off between precision and recall—as precision increases, recall decreases and vice versa—can be balanced by taking the harmonic mean of the two scores. This is the F-measure, given in (74).

(74)

$$\textit{F-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Table 6.1 gives the complete set of experiment results. We experimented with 10 different versions of the topic words and combined topic words/stance-attitude subtrees feature sets, with each version containing topic words with WLM scores $\geq$ a given threshold (the first set contained all topic words with WLM scores ≥ 0, the second set contained all topic words $\geq .1$, and so on). In Table 6.1, we present the highest-scoring version of topic words and combined topic words and stance-attitude subtree features relative to WLM score threshold. In the case of NBM, the best-scoring set of topic words was that set containing words scored $\geq$ .3 while the NBMB model performed best when it used words with WLM scores $\geq .4$. The SVM model performed best for this feature set when the entire set of topic words was used (i.e., WLM score ≥ 0). For the third feature set, which combines both topic words and stance-attitude subtrees, three different WLM thresholds contributed to the best accuracy across all three classifiers. These thresholds were .7, .8, and 0 for the NBM, NBMB, and SVM classifiers, respectively. For each feature set, we also indicate whether the feature set contributed to a statistically significant increase in accuracy relative to the feature set immediately above it (at level $p <$ .05, using McNemar's χ^2 test).

In Figures 6.2 and 6.3 we present the results for these two feature sets for all 10 WLM-score thresholds.

6.5 Discussion

Examining the experiment results in Table 6.1, and figures 6.2 and 6.3, we can make a number of high-level observations. First, all classifiers handily beat both baselines. For each classifier, the combined stance-attitude subtrees and topic words feature set achieved a statistically significant increase in accuracy relative to both baselines at level $p <$ .01, as measured using McNemar's χ^2 test. Second, using accuracy as our performance criterion, the best-performing feature set/classification algorithm combination is an SVM trained on a combination of stance-attitude subtrees and topic words with WLM scores $\geq$ 0. This result was not significantly better than an

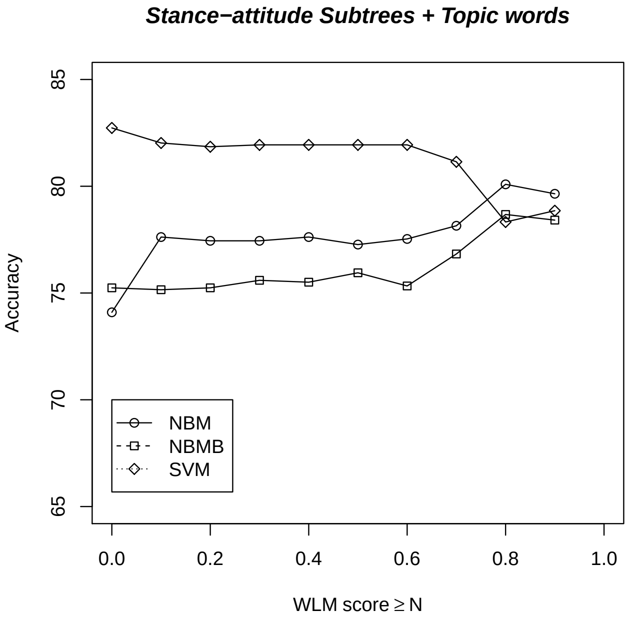


Figure 6.2: Classifier accuracy as a function of increased WLM-score thresholds for all three learning algorithms trained on the combined stance-attitude subtrees and topic words feature set.

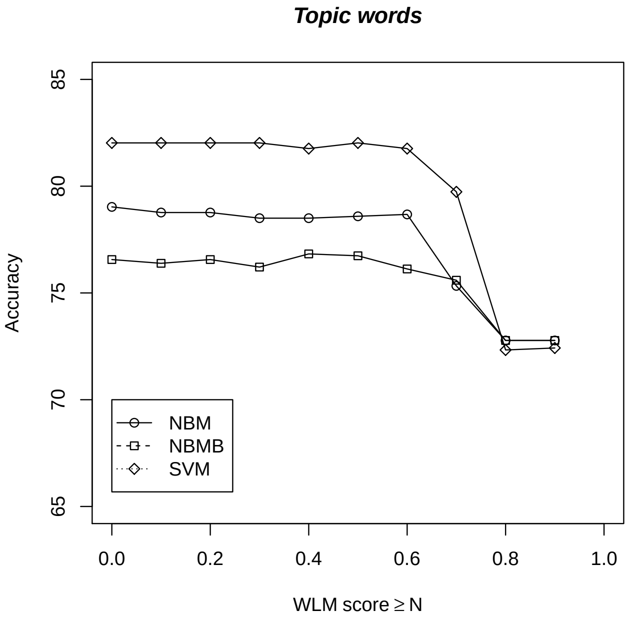


Figure 6.3: Classifier accuracy as a function of increased WLM-score thresholds for all three learning algorithms trained on the topic words feature set.

		for			*against*		
	Acc.	Prec.	Rec.	F	Prec.	Rec.	F
Multinomial Naive Bayes							
Baseline 1: Bag of words	73.1	78.0	63.6	70.1	69.9	82.5	75.7
Baseline 2: S & W features	71.4	71.6	68.7	70.1	71.3	74.1	72.7
St-Att subtrees	79.7	71.4	86.0	78.0	88.2	75.2	81.2
Topic words (WLM score ≥ .3)	80.0	73.3	83.6	78.1	86.4	77.4	81.6
St-Att subtrees + Topic words (WLM score ≥ .7)	**80.4**	73.3	88.5	80.2	88.9	74.0	80.7
Multivariate Bernouli Naive Bayes							
Baseline 1: Bag of words	71.8	71.6	71.8	71.7	72.1	71.8	71.9
Baseline 2: S & W features	73.4	74.1	71.6	72.9	72.9	75.3	74.1
St-Att subtrees	77.4	70.0	85.1	76.8	86.0	71.4	78.0
Topic words (WLM score ≥ 4)	76.7	68.6	86.7	76.6	86.9	68.9	76.9
St-Att subtrees + Topic words (WLM ≥ .8)	**78.8**	70.4	89.4	78.8	89.5	70.6	78.9
SVM (RBF kernel)							
Baseline 1: Bag of words	77.8	75.9	81.4	78.5	80.2	74.4	77.2
Baseline 2: S & W features	73.8	74.1	72.9	73.5	73.6	74.8	74.2
St-Att subtrees	67.3	90.4	28.5	43.4	63.6	97.6	77.0
Topic words (WLM score ≥ 0)	81.8*	79.7	78.7	79.2	83.5	84.3	83.9
St-Att subtrees + Topic words (WLM score ≥ 0)	**82.0**	79.4	79.7	79.6	84.1	83.8	84.0

Table 6.1: Essay-level stance classification experiment results for the highest-scoring feature set/classifier combinations. For each classifier, the highest accuracy model is boldfaced. The highest accuracy model across all three classifiers is boldfaced and underlined. * indicates a statistically significant increase in accuracy (at level $p < .05$, using McNemar's test) relative to the feature set immediately above.

SVM classifier trained on just topic words (WLM score ≥ 0). Third, the overall worst-performing feature set/classification algorithm combination in Figure 6.1 is the SVM algorithm trained on just stance-atttiude subtrees.

In Figure 6.2, which plots classifier accuracy scores against WLM score increments, we can observe a distinctive trend across two of the three classification algorithms — NBM and NBMB— when they are trained on the combined stance-attitude subtrees and topic words feature set. Classifier accuracy increases for both classifiers when words are included with a WLM score ≥ .6. On the other hand, accuracy for the SVM classifier decreases when topic words are included with WLM score ≥ .6. To understand the trend observed in our highest-accuracy model, the SVM classifier, we can examine Table 6.3 which plots the accuracy of the models trained on just topic words relative to WLM thresholds. We observe a clear decrease in classifier performance when higher-threshold WLM scores are included. We can interpret this trend in the following way: Since WLM-scored words $\geq$ .9 tend to have scores approaching 1 and hence are often just lists of the content words contained in each prompt, a classifier trained on words with WLM scores $\geq$ .6 will contain a large number of such words. These include words found in prompts such as *feminist, scienc*, and *technolog* (recall that all topic words were stemmed) which are used in equal measure in both *for* and *against* essays. Thus, they tend to be poor discriminators of stance.

Examining each of these results in more detail, we begin with the results for the NBM classifier. In general, the success of the multinomial model across all feature sets can be ascribed to the general tendency of essay writers to articulate a stance cumulatively by repeating their positions using very similar language. The multinomial model, which calculates its conditional probabilities using frequency counts, is able to capture these redundancies. The authors of the ICLE essays also tend to avoid nuanced, or ambiguous argumentative positions, a tendency that gives rise to highly discriminative features. As expected, many modals, such as the possibility modal *can*, the deontic modal *should*, and the predictive modal *will*, appeared to be good discriminators of stance.

Manual examination of highly discriminative features generated by the NBM model revealed a basic trend: there was often a polarity mismatch between the feature itself and the stance the feature was used to predict. Terms that are discriminative of an *against* stance such as *can-improv, should-forget, should-kill, will-astonish*, and *will-continue*, contain the modals *can, should*, and *will*, which are all found in the *for* section of the stance lexicon. This lends support to the observation made in section 3.4 regarding lexical differences between stancetaking and opinion-bearing language. Unlike opinion polarity, which can be recognized *a priori* as positive or negative (*great* is always positive; *terrible* is always negative), stance polarity must always be evaluated relative to the proposition it is evaluating. For example, the stance word *should* can be used to argue *for* the proposition *Prisoners must be rehabilitated* (e.g., *Prisoners* ***should*** *be reintegrated into society*) but it can also be used to argue *against* that same proposition (e.g., *Prisoners* ***should*** *be punished*).

We also observed that discriminative topic words were not prompt-specific. Consider the topic word *slave*, which is a very good indicator of an *against* stance. We found this word frequently used to argue against propositions contained in the *Feminism* and *Science* prompts, as shown in the sentences given in (75)

(75) a. Housewives are suppressed **slaves** to the men have no will of their own. —*Feminism* prompt

b. The luxury which machines provide for us appeals tremendously to those who come home exhausted after work and are happy to be soothed by their mechanical **slaves**. —*Science* prompt

Similarly, the stemmed aspect word *ideolog* was a very good indicator of *for* stance across the *Degrees, Prisons*, and *Marx* prompts. Since the propositions contained in these prompts contain negative language, a writer aligning herself with this proposition will also make use of negative language (the term *ideology* generally has a negative connotation).

(76) a. If a human creature is surpressed by some sort of useless theory or **ideology** he is susceptible to arouse in himself extremely negative and perverted forces. —*Degrees* prompt

b. In one place the system and the **ideology** which punish faults against life freedom or human dignity can not fall in the same vices; as freedom life or dignity are basic human rights all the people are worthy of that not only those one with a good behaviour. —*Prisons* prompt

c. Obviously the domination of a minority over a majority cannot be made without the support of the suitable **ideology** above all when that majority endures subhuman life conditions. —*Marx* prompt

We next discuss the results of our highest-accuracy classifier, the SVM algorithm. As mentioned in section 6.2, SVMs typically outperform other learning algorithms for text classification tasks, so its high-accuracy performance on our dataset was not entirely surprising. Joachims (2002) explains the superior performance of SVMs for text classification tasks by noting three key properties of these tasks that make them amenable to modeling using SVMs. First, as Joachims empirically demonstrates, natural language text tends to be linearly separable and this property makes it well-suited to decision-boundary learners such as SVMs. Second, SVMs deal well with the large feature spaces typical of text classification tasks. The danger of any classification model learned from such feature spaces is that the learned model will not generalize well. SVMs avoid overfitting by recasting these feature points into a higher dimensional, linearly separable space. Of course, in separating the feature space in this way, there is a risk that a trivial boundary separating the two classes will be generated which, again, can lead to overfitting. In choosing a separating hyperplane from the many (trivial) hyperplanes that linearly separate this feature space, SVMs avoid this problem by choosing the one that maximizes the distance separating the two classes of data—the maximum margin hyperplane.

Third, text classification tasks tend to involve a large number of irrelevant features. In our feature sets, for example, we find such seemingly irrelevant items as *winter* and *some-N-sound* that do not appear to have a role in capturing either prompt topics or the relationship between the writer's stance and the target of that stance. Yet such features often contribute to the performance of classifiers such as SVMs and NB. Joachims (1998), for example, demonstrates that SVMs and NB classifiers trained on the least informative subset of word features (where the informativeness of a given word is measured using the *information gain* criterion of Cover and Thomas, 2012) perform well above chance for text classification tasks. Classifiers that aggressively prune these irrelevant or less informative features will not learn the "dense concepts" (Joachims, 1998) that enable SVMs to learn such high-accuracy models as the essay-level stance classification models reported here.

In Table 6.2 we have given a selection of 15 *for* and *against* term weights used by the classification function returned by the SVM algorithm. This classification function was generated by an SVM trained on the combined aspect word/stance-attitude subtree feature sets, with topic words restricted to those that received WLM scores of ≥ 0.

Examining the list of high weighted *for* terms, we find a number of modals, such as *will* and *can*, and this again seems to confirm the view in the literature that stance is often expressed

for TERM WEIGHTS	
FEATURE	WEIGHT
will-come	0.3322
prioriti	0.3209
do-abl	0.3167
theori	0.3033
convers	0.2743
comput	0.2598
attend	0.2535
affect	0.2482
is-right	0.2455
peopl-help	0.2150
not-interest	0.2131
can-cope	0.2114
it-reason	0.1991
most-import	0.1917
religion	0.1915

against TERM WEIGHTS	
FEATURE	WEIGHT
imagin	-0.2503
noth	-0.2869
more-harm	-0.2286
to-admit	-0.2137
not-interest	-0.2131
therefor	-0.1949
altern	-0.1936
non	-0.1947
limit	-0.1869
caus	-0.1826
can-learn	-0.1789
it-hard	-0.1405
commit	-0.1390
belief	-0.1340
not-provid	-0.1291

Table 6.2: Top 15 *for* and *against* term weights learned by the SVM classifier for the combined stance-attitude subtrees and topic words feature set. Topic words were restricted to those with WLM scores ≥ 0.

modally. High weighted terms in the *against* column are largely non-prompt-specific. The highly weighted term, *not-interested*, for example, is often used by writers when describing people whose existence are presented as evidence against the premises contained in the prompt, as in (77).

(77) a. They are **not interested** in money and success so much. —*Money* prompt

b. The reverse of the medal people who are **not interested** in modern technologies may be more disposed to dreaming because they are not burdened with such specific information but that is not always the case. —*Science* prompt

As with the highly discriminative MNB terms discuseed above, many terms with high *against* weights are drawn from the *for* section of the stance lexicon. These include *imagin, therefor, can-learn*, and *belief.*

While the highest accuracy classifier overall was an SVM, it is worth pausing for a moment to consider how the NBM algorithm was able to achieve an accuracy approaching that of the SVM algorithm. It is at first difficult to discern how the humble NB algorithm, with its simplistic model of natural language can, for some text varieties at least, compete with state-of-the-art classification algorithms such as SVMs and kNN. There is an interesting line of research that provides some insight into the surprisingly effective performance of NB for certain text classification tasks. This research makes three basic points. First, as empirically demonstrated in Domingos and Pazzani (1997), NB is optimal when the training data consists of conditionally independent features. This does appear to be the case with stancetaking text. Second, NB tends to be immune to "concept drift" (Forman, 2006) in text collections. The fact that the terms "President of the United States" and "George W. Bush" co-occur with high-probability in one section of a corpus, while the first of these terms co-occurs with "Barack Obama" with high-probability in another section, has no effect on the classification decisions of a NB learner—again, this is due to the independence assumption. Other, more sophisticated learners, such as the SVM algorithm or kNN, will attempt to capture such idiosyncrasies, leading to lower classification performance. Third, NB is inherently prone to "probability overshoot" when estimating its conditional probabilities. In binary classification settings, this leads to posterior probabilities that are usually very high for one class and very low for its complement class. For this reason, the probability estimates of NB are often poor while its classification decisions are often surprisingly good (Manning et al., 2008, pg. 269).

Chapter 7

The language of supporting arguments

The goal of the next few chapters is to supplement the document-level stance classification model described in chapters 5 and 6 with two additional classification models. The first model identifies a sentence extracted from an ICLE essay as containing a reason given in support of the essay-level stance identified in the last chapter. We call such sentences *supporting arguments* for the essay-level stance. The second model identifies the polarity of a given supporting argument: Does the supporting argument contain a reason in support of a *for* stance or does it contain a reason in support of an *against* stance? Adopting the terminology of Wilson et al. (2009), the first model is termed the *neutral-polar* classification model. It classifies all sentences in a given essay as containing a supporting argument or as not containing such an argument. The second, *polarity classification* model classifies supporting argument sentences as *for* or *against* the essay prompt. Although these models are not built in a two-stage- or cascade-fashion (i.e., the supporting arguments identified by the first model are not used in the second model), they can be used in tandem to identify and classify supporting arguments in stancetaking text. In chapter 10, we provide an example of such an approach.

In this chapter, we lay the theoretical groundwork for the classification features used in our two supporting argument classification models by describing several influential analyses of supporting arguments found in the linguistics and philosophy literature. These analyses are motivated by markedly different research concerns. Grice (1989) and Fraser (2006) are interested in explaining the pragmatic competence underlying a speaker's ability to recognize coherent relations between discourse segments. Toulmin (2004) is primarily interested in explaining the discourse logic underlying practical, or everyday reasoning. Finally, Hyland's (1990) analysis of the rhetorical structure of student argumentative essays has a pedagogical focus. Nevertheless, these analyses have certain common strands that we attempt to capture in the supporting argument classification feature set described in chapter 8. One important point of agreement is that a reader's inference that a given text segment provides a reason for accepting the truth or likelihood of a claim in another segment involves the recognition of certain lexical cues (logical connectives such as *because* and *therefore*). Absent these lexical cues, making this inference involves the recognition of implicit coherence relations holding between two segments. For this reason, analyses of the language of supporting arguments are centered around their relation to the immediate discourse context. This contrasts with analyses of stancetaking language, summarized in chapter 3, which rarely move beyond the sentence boundary.

7.1 Background

In this work, we adopt the colloquialism "supporting argument" (SA) to describe those sentences that contain segments that provide evidence or justification for the reader's acceptance of some other segment containing the writer's central argument. This brief description makes the implicit claim that SAs are always evaluated relative to another, more central argument, and that this central argument must co-occur somewhere in the discourse context. The central argument could be in a nearby clause, as in (78a), or in a nearby sentence, as in (78b). In both examples, the segment containing the SA is boldfaced while the segment containing the argument is italicized.

(78) a. I think yes *the prison system is outdated* because **when a criminal is being imprisoned there are no resources for them to be rehabilitated**.

b. *If he* [i.e., Marx] *had lived nowadays he would have replaced religion by television.* [. . .] **Technology grew and created its own god; it was called television and it was invented for the same purpose as religion: to keep the mass stupid and to make sure that people do not think.**

Since SAs are evaluated relative to another text segment, one often contained in an entirely different sentence, we can describe their linguistic characteristics by drawing on research dealing with meaning beyond the sentence boundary. In particular, we draw on both discourse structure research, which deals with intersentential relationships, and research involving rhetorical and argumentative structure in various text genres. In sections 7.2 and 7.3, we draw on several influential analyses of discourse structure to describe the use of SAs in the ICLE corpus. In these analyses, SAs are generally described in terms of their relationship to another, more central text segment. In the text cohesion account of Halliday and Hasan (1976), for example, these segments express a *causal* relationship between the central argument and the SA, while in the Rhetorical Structure Theory (RST) account of Mann and Thompson (1988) they express *evidence* for the central argument or are designed to *justify* the reader's acceptance of the central argument. In the account of Martin (1992), SAs *enable* or *determine* the reader's acceptance of the central argument. Following Martin, one can view the *consequential* relationship between a central argument and its SA as turns in an exchange, with the SA providing an answer to implicit questions of Why? or How? generated by the central argument, as in (79a) and (79b).

(79) a. I think yes *the prison system is outdated.* (Why?) **When a criminal is being imprisoned there are no resources for them to be rehabilitated**.

b. *If he* [i.e., Marx] *had lived nowadays he would have replaced religion by television.* (Why?) **Technology grew and created its own god; it was called television and it was invented for the same purpose as religion: to keep the mass stupid and to make sure that people do not think.**

While these accounts differ somewhat in their points of emphasis, there is agreement regarding two basic aspects of SAs. First, SAs bear an asymmetric relationship with a more central text segment. In the RST account of Mann and Thompson (1988), for example, SAs serve as *satellites* to the central or *nuclear* argument, while in Grice's (1989) speech act-theoretic account, SAs communicate information regarding *non-central* or *higher-level* speech acts. These speech acts serve to contribute to the meaning of *central* or *ground-floor* speech acts communicated by segments such as the central argument in (78a) *the prison system is outdated.* Fraser (1990), in his discussion of discourse markers that often introduce SAs, such as *hence, moreover*, and *therefore*, describes these markers as signaling speaker *comment* on a segment in the prior discourse. In later work, Fraser (2006) describes SAs as one part of a semantic relationship of *inference* holding between two segments.

A second point of agreement involves the structural characteristic of SAs. The relationship between an SA and the central argument it is meant to justify is inter-sentential or inter-clausal. This observation is borne out by the frequent use of two-place discourse markers such as the subordinators *because* and *since*, which link clauses, as in (78a), and sentence- or clause-initial markers such as *moreover*, *consequently*, and *thus*, which introduce clauses or sentences, as in (80) and (81). In these examples, arguments indicative of the writer's global stance are italicized and SAs are boldfaced. Markers introducing SAs are underlined.

(80) *The abolition of the compulsory military service is a current project in 1992 with which I completely agree as I am convinced that an army consisting of professional soldiers would be more efficient.* [. . .] Moreover **the barracks are uncomfortable the food is not very good, they have to wake up very early...**

(81) There are quite a few people who think that *higher education is unnecessary* and that *there is no point in studying at university*. [. . .] And thus **most of these graduates become frustrated due to the lack of proper occupation** and consequently **much energy talent and knowledge is wasted**.

Theoretical accounts of discourse markers such as *moreover* and *consequently* in (80) and (81), respectively, note their role in encoding "procedural meaning" (Fraser, 1990) by relating segments of discourse. This idea has its precursor in Halliday and Hasan's (1976) description of discourse markers such as *because* as "conjunctive devices" signaling a cohesive tie between two text segments. Aside from their value in theoretical accounts of discourse relationships, discourse markers play a key role in the construction of NLP knowledge sources such as the RST corpus of Carlson et al. (2003) and the Penn Discourse Treebank (Mitsakaki et al., 2004; Prasad et al., 2008). Both of these resources were partially constructed using lists of discourse markers as clues to semantic relationships between text segments. In section 7.2.1, we describe in more detail the role of discourse markers in signaling SAs in the ICLE corpus.

Along with discourse structure accounts of SAs, many researchers have analyzed SAs within frameworks that stress their importance as rhetorical or argumentative devices. As with the language of stancetaking, interest among philosophers and logicians in the structure of argumentation is rooted in classical antiquity, in particular Aristotle, whose distinction between *apodictic*, *dialectical*, and *rhetorical* arguments is still maintained in the contemporary division between analytic arguments, or arguments that are true by virtue of their meaning, *critical discussion*, or arguments designed to resolve competing views of several speakers, and so-called *monological* arguments, or arguments designed to convince or persuade an audience of the truth of a claim. Two influential contemporary models of critical discussion are the argument schemes of Walton (1996) for *presumptive reasoning*, and the *pragma-dialectical* model of Van Eemeren and Grootendorst (2004). Presumptive, or plausible reasoning, is the reasoning of everyday conversation and inquiry (and, arguably, legal discourse), and involves generalizing from less than conclusive grounds to a conclusion that is at least presumptively valid. It is a dialectical style of reasoning in the sense that the presumption on which the conclusion rests can be challenged by an interlocutor by drawing on a series of "critical questions" defined by Walton. The pragma-dialectical model of critical discussion is also presented as a model of non-deductive reasoning and defines a set of rules presumed by the conversation participants during the process of resolving a difference of opinion.

Rhetorical, or monological reasoning is the style of argumentation that is most relevant to the analysis of SAs as they occur in argumentative essays. This form of argumentation involves a single speaker or writer using “available means of persuasion” (Hohmann, 2002, pg. 43) to persuade a specific audience of the truth of a claim. The available means in question are determined by the discourse context of the argument—in particular, the targeted audience—and the speaker’s use of “intrinsic” means, such as verbal dexterity or inventiveness, or “extrinsic” means, such as laws and documents. This basic conception of rhetorical argumentation informs much contemporary research involving the rhetorical structure of written argumentation. Our analysis of the rhetorical structure of SAs in section 7.3 will draw on two influential models of written argumentation: Toulmin’s (1958) model of argument patterns as they occur in contexts involving everyday, or “practical reasoning” contexts such as courtrooms and conversations, and the rhetorical “move” analysis of Swales (1990) adapted by Hyland (1990).

7.2 Lexical features

As we saw in chapter 3, research on the lexical aspects of stancetaking language has generally focused on the use of modal and evidential language such as *ought to*, *should*, *possibly*, and *definitely*, which are used to endorse or disclaim an accompanying proposition. The typical ICLE essay will often contain a first paragraph sentence indicating the writer’s overall stance toward a given prompt using such language, as in (22b), repeated below.

(22b) Something **ought to** be done to improve the prison system and the way to rehabilitate criminals **should** be found

Given the standpoint in (22b), the goal of an SA (and the goal of argumentation itself) is “aimed at convincing a listener or reader of the acceptability of the standpoint” (Van Eemeren and Grootendorst, 2004, pg. 2). If we imagine the writer and reader as participants in an exchange, a good SA for (22b) will convincingly answer the reader’s question,“Why should something be done to improve the prison system, etc.?" Reading further in the essay from which (22b) was extracted, we find the SA in (82b) offered in response to this implied question. Together, (22b) (reproduced below as (82a)) and (82b) form a coherent argumentative pair, as given in (82).

(82) a. Something ought to be done to improve the prison system and the way to rehabilitate criminals should be found.
b. Maybe if the society gave its criminals a real opportunity of rehabilitation, less crimes would be committed.

The SA given in (82b) is located several sentences away from (82a) and is not explicitly marked as an SA using any of the *inferential markers* identified in Fraser (2006), such as *because* or *since*. How then do readers make the inference that (82b) is offered in support of (82a)? In *coherence*-based theories of discourse connectedness (Hobbs, 1979; Mann and Thompson, 1988; Renkema, 2004), implicit *coherence relations* allow readers to recognize that two text spans are related hypotactically. In the case of (82), a coherence relation of *evidence* is presumed to hold between these two text spans—without this presumption, (82) would be considered a non-sequitur.

7.2.1 Discourse markers

The recognition that (82a) and (82b) form an argumentative pair is one component of a more general pragmatic competence that allows readers to recognize both coherent discourse, as exemplified by (82), and incoherent discourse, as in the sentence pair given in (83).

(83) a. Something ought to be done to improve the prison system and the way to rehabilitate criminals should be found
b. I like plums.

One way of explaining the coherence of (82) and the contrasting incoherence of (83) along Gricean lines is to invoke Sperber and Wilson's (1986) principle of *Optimal Relevance*. This principle, it is claimed, governs every act of communication. Briefly described, the principle guarantees, 1) that the utterance in question is worth the processing cost incurred by the hearer/reader and, 2) that the presentation style of the utterance is the most relevant that the speaker/writer is capable of. In the case of (83), the second guarantee has been violated. Additionally, Sperber and Wilson (1986) use this principle to explain the existence of markers, such as *so, since*, and *because*, that explicitly guide the reader's understanding of the discourse relationship between two segments. Explicit use of a marker relating two segments can be viewed as an effort to maximize the relevance of the second segment while minimizing processing costs. This strategy is employed in cases of interpretative ambiguity, where use of explicit relational markers is intended to guide the reader to the correct inference. A well-known example is the sentence pair given in Hobbs (1979, pg. 78, ex.3), which is adapted below as (84).

(84) a. John can open Bill's safe.

b. He knows the combination.

There are two interpretations of the pair given in (84). If (84b) is functioning as a premise, then it is interpreted as evidence that John can open Bill's safe. On the other hand, if it is functioning as a conclusion, then it is interpreted as a consequence of the assumption that John can open Bill's safe. The markers *after all* and *so* can resolve this ambiguity, with *after all* marking (84a) as a premise and *so* marking it as a conclusion. This is shown in (85).

(85) a. John can open Bill's safe.

b. After all (premise)/So (conclusion), he knows the combination.

In Fraser's (1990) taxonomy of *pragmatic markers*, *discourse markers* (DM) such as *after all* and *so* are kept separate from *commentary* markers such as *frankly* and *allegedly* which comment on the accompanying segment and hence add to the interpretation of that segment. Instead, the role of DMs is to encode *procedural meaning*: they serve as signposts guiding the reader's inference that the current segment is related somehow to a segment in the prior discourse, but they do not themselves affect the truth conditions of either of the related segments. Fraser's notion that DMs are non-truth-conditional can also be found in Grice's speech act-theoretic (1989) account of DMs. For Grice, use of DMs such as *moreover* signals a non-central "speech act of adding, the performance of which would require the performance of one or another of the central speech-acts" somewhere in the prior discourse (Grice, 1989, pg. 122). Since they signal the performance of a speech act (in the case of *moreover*, a speech act of *adding*; in the case of *but*, a speech act of *contrasting*), they serve to indicate conventional implicatures. Thus, they do not directly affect the truth conditions of their accompanying utterances. For example, the use of *so* in (85b), signals that a non-central speech act of *explanation* holds between (85a) and (85b); however, taken individually, the truth conditions of *John can open Bill's safe* and *He knows the combination* remain unchanged.

Fraser (2006) presents a taxonomy of numerous classes of DMs, each encoding different semantic relations between text segments. In the present context, we are mainly interested in those DMs that encode the semantic relations that typically involve SAs in argumentative essays. Since argumentative DMs are massively polysemous (Prasad et al., 2008, for example, identify twelve distinct senses of *while* and nine distinct senses of *if*), it is not possible to straightforwardly identify the distribution of occurrence of argumentative DMs in the ICLE corpus. However, a manual examination of the ICLE corpus reveals that a small class of commonly used argumentative DMs frequently appear in SAs. In the examples given in Table 7.1, we adopt the theory-neutral terminology of the Penn Discourse Treebank (PDTB), a resource that we will discuss in more detail in chapter 8.3.1. The mappings of DMs to PDTB relations in these examples are based on the descriptions of explicit discourse relations captured by the PDTB as presented in Prasad et al. (2007). All supporting argument examples in Table 7.1 were identified by annotators in the supporting argument annotation task described in 9.1.

While Sperber and Wilson's (1986) analysis of the use of explicit DMs is processing-oriented and Grice's (1989) analysis is speech act-oriented, there are other, genre-specific reasons that we

PDTB SENSE	DISCOURSE MARKERS	SUPPORTING ARGUMENT EXAMPLES
Conjunction	additionally, also, as well, besides, even then, further, furthermore, in addition, in fact, plus, similarly, specifically, ultimately	a. **Additionally** women frequently take advantage of their right to get numerous sick leaves either because of children's health problems or their own's. b. Swallowing every day ready-made-images can **also** lead to the habit of giving in to your own dreamworld.
Contrast	although, but, by comparison, even though, however, nevertheless, on the one hand...on the other hand, rather, whereas, while	a. It could be argued **however** that it is exactly imagination and dreams which incite creativity. b. But from the end of WW2 until the end of the Cold War the importance of each nation's military forces seems to have gradually decreased **while** the importance of international organizations like the UN and NATO have increased.
Hypothetical	as long as, especially if, even if, if, if and when, if...then, only if, only when, particularly if, unless	a. **If** both men and women are thought as people without considering sex difference the inequality between them would not occur. b. **As long as** we have art there is still place for dreams and fantasy.
List	also, finally, moreover, separately, then	a. And **finally** isn't imagination more important than knowledge as Einstein said? b. **Moreover**, the idea of compulsory military service is out of date.
Opposition	although, by contrast, in contrast, even as, even though, meanwhile, neither...nor, on the contrary, on the other hand, though, yet	a. **Although** television is the distraction of the masses it also brings them dreams imagination and culture. b. **In contrast** absence of this can affect a psychic balance of man in a negative sense.
Reason	apparently because, because, especially as, especially because, insofar as, just because, mainly because, now that, particularly as, since, so	a. I don't agree with the topic **because** I think there are a lot of feminist who are very reasonable, as for example Virginia Woolf. b. **Since** contemporary educational system is not sufficient and does not fulfil the needs of real life it should be changed.

Table 7.1: Discourse markers commonly used to indicate supporting arguments in the ICLE corpus, along with PDTB senses and examples.

might find comparatively frequent use of DMs in a typical ICLE essay. The first reason involves the stylistic characteristics of academic prose sub-genres such as student essays. Genres and registers, by the definitional criteria given in Biber and Conrad (2009), “differ in their characteristic distributions of pervasive linguistic features, not in the single occurrence of an individual feature” (Biber and Conrad, 2009, pg. 9). A reader’s ability to recognize genre distinctions can be partly ascribed to the characteristic distribution of certain classes of words. For instance, recognizing that a given text is an instance of the academic prose genre, while another is an instance of the newspaper editorial genre is often a matter of noting the frequency of occurrence of DMs such as *neither...nor* or the use of DMs such as *nevertheless* in sentence-initial or non-sentence-initial position. Biber et al. (1999) conducted a study of the distribution of mostly non-polysemous, simple coordinators such as *nevertheless* and *neither* and correlative coordinators such as *both...and* and *neither...nor* across conversation, newswire, fiction, and academic corpora and found significant distributional differences. Correlative coordinators, for example, are most frequently used in academic prose and are almost never found in conversation and newswire text (Biber et al., 1999, pg. 84).

A second resason we might find comparatively frequent use of DMs in the ICLE corpus involves the English writing competence of the authors of these essays. The essays contained in the ICLE corpus are written by ESL students representing 16 language backgrounds. Several studies dealing with the comparative use of DMs in academic essays written by English L1 and English L2 speakers show that, in general, English L2 writers tend to overuse DMs. Overuse of additive DMs such as *moreover* and *besides*, contrastive DMs such as *but* and *however*, and result DMs such as *therefore* have been reported in quantitative studies of DM usage in ESL writing (Granger and Tyson, 1996; Yang and Sun, 2012; Lee, 2013). Granger and Tyson (1996), using quantitative results from an early version of the ICLE corpus, analyzed the overuse of DMs in French L1/English L2 student writing as a transfer effect: Use of DMs is significantly more frequent in French writing than in English writing. Lee (2013), in his analysis of Korean L1/English L2 academic writing, ascribed the overuse of exemplifiers such as *in the case of* and enumeratives such as *first, second*, etc. to the test-driven culture of English writing instruction in Korea, which emphasizes the use of explicit markers of argumentation in written English test responses.

7.3 The rhetorical structure of argumentative essays

In this section, we consider two models of rhetorical structure in student argumentative essays. The first model considered is the influential Toulmin model (Toulmin, 2004), which deals with the internal structure of individual arguments. We will also examine Hyland's (1990) rhetorical move analysis of student argumentative essays which captures the global rhetorical structure of this text genre. We find that these two models are most useful to the analysis of supporting arguments in ICLE essays when they are applied at different levels of description. The Toulmin model is best used to describe the structure of individual arguments found within text units such as paragraphs or sentence pairs, and the Hyland model is helpful as a description of the rhetorical relationships between those text units across the entire essay.

7.3.1 Toulmin's (1958) model of argumentation

Toulmin's (1958) model of argumentation was initially developed as a critique of formal models of argumentation and, more generally, of the idea that any valid argument can be defined in rigidly deductive terms (Toulmin, 2004, pg. vii). For Toulmin, the vast majority of arguments, as they occur in everyday conversation or in courtrooms, cannot be deductively defined and rely instead on an informal schemes of practical reasoning. Toulmin's model can be defined as a group six interrelated elements. These six elements, in turn, can be defined as two triads. The first triad involves a *Claim*, which is the proposition that is being argued for, *Grounds*, which is the proof or evidence offered in support of the *Claim*, and the *Warrant*, which is the inference linking the *Claim* with the *Grounds* (Toulmin, 2004, pg. 92). The *Warrant* is typically unstated and relies for its effect on the presumed background knowledge shared by the speaker/writer and hearer/reader. Many intersentential argument patterns observed in the typical ICLE essay can be explained using the Toulmin model. Examples of *Claim-Grounds* pairs are given in (86).

(86) a. The fact is that we have ceased to be spiritual and simple beings and have turned into too practical down-to-earth people.[Grounds] That is why there is no longer place for imagination and dreaming in our present life. [Claim] —*Science* prompt
b. Where a lot of money is, evil exists.[Claim] Money determines social position in society... [Grounds] —*Money* prompt
c. All armies should consist entirely of professional soldiers: there is no value in a system of military service.[Claim] This is a self-evident statement. Military service is not something you enjoy or do for pleasure; it's something you are obliged to do. [Grounds] —*Armies* prompt

In (86a), the writer's *Grounds* for supporting the *Claim* that there is no longer a place for dreaming and imagination is the lack of spirituality and simplicity in modern life. The implied *Warrant* connecting claim to grounds is *Practical-minded people lack imagination*. The writer's grounds for the *Claim* in (86b) is the prevalent correlation between social status and wealth. Here, the implied *Warrant* is *It is unjust (evil) that wealth should determine social position*. Finally, in (86c), the writer's observation that compulsory military service is not done out of pleasure, but out of obligation, implies the *Warrant*, *Activities done out of obligation are of no value*.

The second triad of Toulmin's model involves a *Qualifier* which calibrates the speaker/writer's degree of commitment to the *Claim*, a *Backing* which provides evidence for the *Warrant*, and a *Rebuttal* which states any restrictions that might apply to the *Claim* (Toulmin, 2004, pg. 94). The category of *Qualifier* will not be discussed here since many of the items associated with this category were already described in our discussion of evidential markers in section 3.6. The essay from which (86a) was extracted provides us with several examples of *Backing* sentences, one of which is given below as (87), along with its accompanying *Claim*, *Grounds*, and *Warrant*.

(87) That is why there is no longer place for imagination and dreaming in our present life.

> [Claim] The fact is that we have ceased to be spiritual and simple beings and have turned into too practical down-to-earth people. [Grounds] *Practical-minded people lack imagination*. [Warrant] People no longer care for reading and intellectual talks because in our speed-obsessed world they do not have the time and desire to spend hours with them. [Backing]

Toulmin's final category, *Rebuttal*, captures those parts of arguments that apply restrictions or note exceptions to the validity of the *Claim*. The author of (86a), for example, after declaring that there is no longer room for imagination in contemporary life, notes the existence of alternative communities that have renounced technology and science, as given in (88a). The author of the essay from which (86b) was extracted, after listing various evils resulting from the pursuit of wealth, concedes in (88b) that money is somewhat beneficial.

(88) a. More and more people are beginning to ask the question Do we really need all these technological achievements to live happily? And more and more of them answer it negatively.Those people educate their children to hate commercialism and to respect their relationship with the people they encourage their children to live with animals in order to be closer to Nature.

b. Money alone does not make us happy but it contributes a little to our welfare.

In general, the Toulmin categories that are most relevant to the task of supporting argument classification are *Grounds* and *Backing*. Many sentences identified as supporting arguments in the annotation task described in section 9.1 can be analyzed as *Grounds* or *Backing* using Toulmin's scheme. We can illustrate the association between sentences identified by annotators as supporting arguments and particular Toulmin categories by examining multi-sentence argument chains extracted from essays annotated for both essay-level stance and sentence-level supporting argument stance. Additionally, drawing on our discussion of DMs in section 7.2.1, we can discover possible lexical cues to *Grounds* or *Backing* sentences by looking for any occurrences of DMs in such sentences. In Table 7.2, we provide analyses of two complete *Claim*-*Grounds*-*Backing* chains. For each sentence in the chain, we provide the Toulmin category and supporting argument tag. Any DMs identified in a sentence are also given and are categorized using the PDTB scheme.

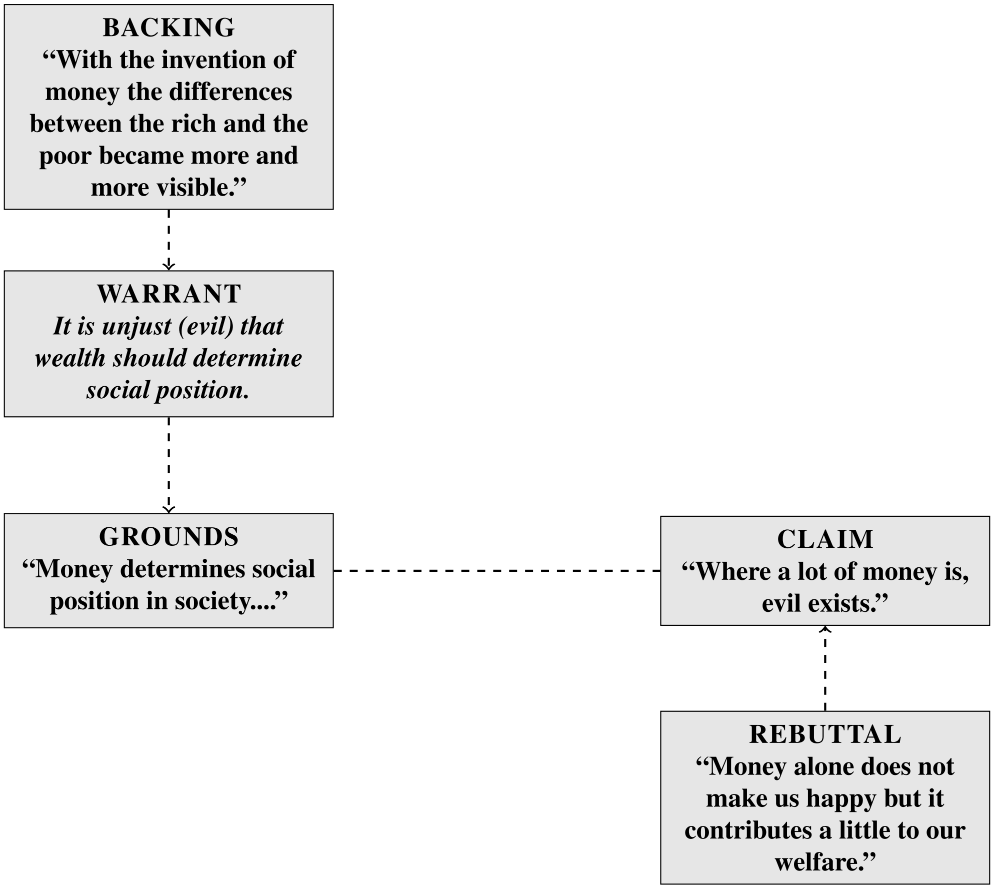

BACKING
"With the invention of money the differences between the rich and the poor became more and more visible."
WARRANT
It is unjust (evil) that wealth should determine social position.
GROUNDS
"Money determines social position in society...."
CLAIM
"Where a lot of money is, evil exists."
REBUTTAL
"Money alone does not make us happy but it contributes a little to our welfare."

Figure 7.1: A complete Toulmin argument chain taking a *for* stance toward the proposition, "Money is the root of all evil."

PROMPT/ ESSAY STANCE	SENTENCE	TOULMIN CAT.	SUP ARG?	DMS (PDTB SENSE)
Money/*for*	Where a lot of money is, evil exists.	*Claim*	No	n/a
	Money determines social position in society...	*Grounds*	Yes	n/a
	With the invention of money the differences between the rich and the poor became more and more visible.	*Backing*	Yes	n/a
	In effect **if** one had a possession, he/she had to have **also** the right amount of money to maintain it and to pay taxes.	*Backing*	No	*if* (hypothetical), *also* (conjunction)
	So, all around the world, there were a little number of rich and a huge number of poor and destitutes.	*Backing*	Yes	*So* (reason)
Armies/*for*	Why is it necessary to base the protection of a country in a compulsory military service when it could be done by a professional and well prepared army?	*Claim*	No	n/a
	One more thing to add is the importance of the interest with which ones (the professional army soldiers) and the others (the military service soldiers) do their duty.	*Grounds*	No	*One more thing to add* (list)
	As military service is compulsory, young soldiers do not take any kind of interest in what they are doing.	*Backing*	Yes	*As* (reason)
	On the other hand professional soldiers have no option but to take a great interest in it because in the end it is their job what they are doing and they are being payed for it.	*Backing*	Yes	*On the other hand* (contrast)

Table 7.2: *Claim-Grounds-Backing* chains extracted from two ICLE essays. Both essays were annotated for document-level stance and sentence-level supporting argument stance. The Toulmin category for each sentence in the argument chain is provided, along with its supporting argument/not-supporting argument tag and any identified DMs.

7.3.2 Hyland's (1990) model of rhetorical organization in student argumentative essays

Text genres are traditionally defined as complete texts featuring specialized expressions and a predictable rhetorical organization, both of which "conform to the culturally expected way of constructing texts belonging to the variety" (Biber and Conrad, 2009, pg. 16). Examples include business letters, newspaper articles, and memos. Swales (1990) examined expert and student prose in academic settings from a genre perspective and identified the rhetorical "moves" of research genres such as Ph.D. dissertations (Swales, 2004), research articles (Swales, 1990; Swales and Najjar, 1987), and various other examples of genres typically found in academic settings such as application letters, recommendation letters, and grant proposals. The move analysis of text genres has been applied in the context of corpus-based discourse analysis (Biber, 2007) and has been adopted by fields with a pedagogical focus, such as language and science education (Hyland, 1990; Hyland and Milton, 1997; Jordan, 1997). Areas of composition studies that involve the comparative analysis of L1 and L2 academic writing (Hyland and Milton, 1997) have also adopted the move analysis framework.

Hyland's (1990) move analysis of the student argumentative essay genre is functionally oriented: textual units are defined in terms of their communicative purpose rather than their content. The communicative purpose of the argumentative essay is to persuade the reader of the validity of a proposition. In the argumentative essay, this purpose is realized by the three familiar stages of *Thesis, Argument*, and *Conclusion*. Within each of these stages, the rhetorical progression of an argument or set of arguments is realized by a set of both obligatory and optional moves. Table 7.3 (adapted from Hyland, 1990, Table 2) contains each of Hyland's stages along with a sentence from an ICLE essay exemplifying that stage. As shown in this table, sentences drawn from ICLE essays responding to the *Science* prompt can be neatly mapped to each of Hyland's stages.

Table 7.3: Hyland's (1990) move analysis of the student argumentative essay. Optional moves are bracketed.

Stage	**Move**	**Example**
Thesis	*Proposition*. The proposition to be argued.	"I agree that the prison system should be reformed."
	[*Gambit*]. Attention grabber, controversial statement, dramatic illustration.	"Walking down the street in a slow pace, I try hard to keep these dreamy thoughts of mine alive."
	[*Information*]. Background material; contextualizes the essay topic.	"We live in a world which is, to a great extent, run by machines."
	[*Evaluation*]. Brief support of proposition.	"Imagination is an inseparable part of human character...."
	[*Marker*]. Introduces a list.	"The Women's Liberation Movement aims at ending sexual discrimination and it serves this purpose well for four reasons..."
Argument	[*Marker*]. Signals introduction of claim.	"First, one of the main reasons..."
	Restatement. Repetition of proposition.	"There is always place for dreaming and imagination in our modern world."
	Claim. Reason for acceptance of proposition. Can be based on (a) shared assumptions between writer and reader, (b) generalization based on evidence, (c) or, force of conviction.	"While I cannot deny that, I still hold a view that it is exactly dreaming and imagination which activate all science and technology too."
	Support. Gives grounds supporting the claim. Often involves (a) describing the assumptions underlying the claim, (b) or, providing evidence or references.	"And by thinking of just a few inventors and scientific geniuses, like Einstein, for example, can't we describe them as the greatest dreamers and people with the most vivid imagination ever?"
Conclusion	[*Marker*]. Signals conclusion boundary.	"In conclusion..."
	Consolidation. States significance of arguments to the proposition.	"As shown above, this cycle functions perfectly in our century: The modern scientific, technical and industrial development encourages us to dream and strive to have our dreams fulfilled (in simulative or real ways)."
	[*Affirmation*]. Restates proposition.	"To sum up I think that dreaming and imagination was, is and will always be a part of humanity."
	[*Close*]. Extends significance of proposition.	"We must ask ourselves what will become of us in the future if our lives become completely dependant on machines."

7.4 Summary

We began this chapter by noting a key difference between the language of stancetaking and the language of SAs: while the stance of a document or sentence can be captured by intra-sentential features such as lexical cues and inter-clausal relations, SAs are best analyzed as a discourse-level phenomenon. For readers, the inference that a given text segment in an argumentative essay is an SA for a claim made in another text segment involves two steps. First the location of the SA in the essay must be noted and, second, the (explicit and implicit) discourse cues linking the claim and the SA must be identified.

In section 7.2, we described the lexical features of SAs, adopting Fraser's (2006) term *discourse markers* to describe lexical cues, such as *because*, *therefore*, and *although*, that guide the reader's inference linking an SA to a central argument. For Sperber and Wilson (1986), the existence of this option in natural language can be explained by a principle of Optimal Relevance, which is mutually presumed by speaker and hearer. In light of this principle, the recourse to explicit discourse markers can be explained as the speaker's attempt to maximize the relevance of an utterance while minimizing the processing costs incurred by the hearer. In Grice (1989) and Fraser's (2006) speech act-theoretic account, discourse markers serve as signposts guiding the reader's inference that two text segments are linked somehow, but they do not change the truth conditions of those two segments. Additionally, we suggested in section 7.2.1 that the frequent use of discourse markers in the ICLE corpus could be ascribed to the genre conventions of the argumentative essay, and to the student authors' English language competence.

In section 7.3, we described two models of rhetorical organization, the Toulmin (2004) model, which describes the logical structure of informal, or practical reasoning, and Hyland's 1990 rhetorical "move" model of argumentative text organization. We found that the Toulmin categories most relevant to an understanding of SAs are the categories of *Grounds* and *Backing*: the former category describes those text segments that serve to provide proof or evidence of a *Claim*, while the latter provides evidence for an implicit argument (the *Warrant*) linking the *Grounds* to the *Claim*. The second model of rhetorical organization considered in this section was Hyland's model of rhetorical organization in student argumentative essays. We found that many of the organizational elements of the typical ICLE essay could be neatly mapped to each of Hyland's categories.

Chapter 8

A feature scheme for supporting argument classification

In this chapter, we describe the construction of a feature scheme designed to capture the linguistic characteristics of SAs described in the last chapter. These features incorporate the lexical, discourse structure, and rhetorical structure characteristics of SAs described in the last chapter and are used in the supporting argument classification experiments described in chapter 9. We present features for two distinct classification models. The first, *neutral-polar* classification model identifies any supporting arguments in an essay. The second, *polarity* classification model classifies SAs as providing a reason *for* or *against* a claim.

8.1 Related work

There is a small, but growing body of research in NLP dealing with the identification and classification of reasons given in support of document sentiment. Kim and Hovy (2006) introduced *opinion reason identification* as a sub-task of Sentiment Analysis and presented a system designed to extract single-sentence *pro* and *con* opinion reasons from a corpus of product, company, and service reviews extracted from the online review sites *epinions.com*[1] and *complaints.com*.[2] Since there exist no corpora annotated for opinion reasons, Kim and Hovy (2006) employed a novel corpus collection procedure in order to generate a sizable training corpus of reviews containing sentences annotated as *pro*, *con*, or *neither*. Both *epinions.com* and *complaints.com* invite review authors to provide *pros* and *cons* along with their reviews. These are generally short, succinct phrases such as *beautiful display* and *not something you want to drop*. Given a review and its accompanying author-provided *pro* and *con* phrases, all sentences in the review are automatically labeled *pro*, *con*, or *neither* by finding sentences that maximally overlap with the accompanying *pro* or *con* phrase and then assigning that phrase's *pro* or *con* tag to each of those sentences. Any sentences that did not exhibit overlap with a *pro* or *con* phrase are tagged as *neither*. This process resulted in a training corpus of ∼300,000 sentences tagged as *pro*, *con*, or *neither*.

Kim & Hovy trained a two-stage maximum entropy classifier on their set of ∼300,000 sentences. The first-stage classifier was designed to identify any sentences as opinion-bearing and the second-stage classifier was designed to classify these stancetaking sentences as either *pro* or *con*. Standard text classification and Sentiment Analysis features were used: ngrams, sentence position, and a lexicon of opinion-bearing words generated from WordNet synsets and newswire text containing subjective content (letters, editorials). Test data included reviews extracted from *complaints.com* matched to the topics included in the training data, and a set of 18 reviews containing sentences manually annotated as *pro* reasons, *con* reasons, or as *neither*. The highest accuracy overall for the identification of opinion reasons task was 76.2% (relative to a 57.7% baseline) using only ngram features and the subset of the test data dealing with mp3 player reviews. The highest accuracy for the opinion reason polarity task was 57.1% (relative to a 50.7% baseline), again using only ngram feature with test data restricted to reviews dealing with mp3 players.

Subsequent research in opinion reason mining also made use of product review corpora. Brana-

[1] `http://epinions.com`

[2] `http://complaints.com`

van et al. (2009) used free-text keyphrases generated by review authors to identify particular *properties* of reviews, such as “good food,” “good service,” etc. Such properties are roughly comparable to Kim and Hovy’s *opinion reasons*. Zirn et al. (2011), using multiple lexicons and discourse relationship-based features, present a system that labels discourse segments as positive or negative. Many of these segments include opinion reasons similar to thiose identified in Kim and Hovy, such as *because the quality is bad*. Finally, in Sauper et al. (2011), a system is presented that first identifies properties of products and services reviewed in social media review snippets and then identifies the opinion polarity of these properties. The properties identified and classified in Sauper et al. are comparable to opinion reasons: In *It’s the best sushi I’ve ever had*, for example, the property *best sushi* is a reason given in support of a positive review of a restaurant.

The work most similar to the SA identification and classification research presented here is Arora et al. (2009) and Kwon et al. (2006). Arora et al. (2009) describe a new sentence-level classification task involving the automated categorization of product review claims as either *bald claims* or *qualified claims*. Bald claims (a.k.a., *bare assertions* or *ipse dixit* assertions) are claims presented without evidence or qualification such as *Nikon has terrible customer service*. By contrast, qualified claims present evidence for a particular claim (though, as presented in Arora et al., 2009, the claim itself might not be part of the qualified claim) such as *Took them* [i.e., Nikon] *a whole 6 weeks to diagnose the problem*. As described, the system of Arora et al. resembles an opinion reason mining task with qualified claims comparable to opinion reasons. Approximately 1200 review sentences were manually annotated as bald or qualified claims. Ngrams and shallow syntactic features such as $\langle part\_of\_speech, word \rangle$ tuples were used to train an SVM, resulting in an accuracy of 72.6% (relative to a bag-of-words baseline of 70.6%) for qualified claim classification, and 36.0% (relative to a 31.0% bag-of-words baseline) for bald claim classification.

An example of opinion reason mining outside of the product review domain is presented in Kwon et al. (2006). The research of Kwon et al. can be considered an example of *argument reason* mining and is therefore closest in spirit to our task of SA mining. The system in Kwon et al. is presented in the context of the *eRulemaking* (electronic rulemaking) phenomenon, or the use of digital technologies, such as email and online message forums, to debate and influence the crafting of governmental regulations and rulemaking. The goal of NLP research involving eRulemaking data is the classification and summarization of the opinions and arguments contained in these data. Summarizing the massive number of opinion and arguments related to a particular rule or regulation could be of some benefit to rule-writers who require at-a-glance assessments of the more prevalent arguments that favor or oppose a particular rule. Kwon et al. present a system that identifies key aspects of arguments supporting or opposing regulation proposed by the Environmental Protection Agency. The system was trained on email data annotated for both argument structure (a hierarchical scheme that captures inter-sentential or -clausal relationships between a claim and reasons supporting or opposing that claim) and "subtopics" (topics that are typically invoked in the context of debate regarding environmental regulation such as government responsibility, health concerns, and economic costs). The performance of each component of the system was evaluated separately, with the opinion classification component achieving an accuracy of .77 and the argument structure identification component achieving F-scores of .55 and .65.

Drawing on research in both aspect-based SA (cf. section 2.4.1) and automatic summarization, Glaser and Schütze (2012) introduced a new single-sentence summarization task for sentiment-bearing text such as product reviews. This task involves extracting a single *supporting sentence* from a given review that includes the overall opinion polarity of the review along with a reason for that opinion. A suggested use-case for single-sentence summaries involves an end-user who would like an at-a-glance assessment of current sentiment regarding a company or product plus a reason for that sentiment. In Glaser and Schütze's system, the generation of single-sentence summaries of reviews occurs in the following manner. All sentiment-bearing sentences in the review were first identified using a Maximum Entropy Classifier trained on a bag-of-words representation of positive and negative polarity review text. Reasoning that supporting sentences are generally expressed using topic-specific noun phrases such as *the zoom, the video*, and *the colors*, Glaser and Schütze identified the *n* sentiment-bearing sentences in a review that are likely supporting sentences by means of a weighting scheme that sums the absolute and relative frequencies of all nominal elements in each sentiment-bearing sentence. The resulting sum serves as that sentence's supporting sentence score and the top-scoring supporting sentence is considered the best single-sentence summary of the review.

Glaser and Schütze introduced a novel, crowdsourced evaluation method for their single-sentence summarization system. Developing gold-standard data for a document sentiment summarization system is time-consuming and costly. The task of reading an entire document and selecting the *n* sentences that both summarize the sentiment associated with that document and also give a good reason in support of that sentiment requires significant training and is likely too complex to be completed by crowdsourced annotators. Rather than asking their crowdsourced annotators to read each review in its entirety and to extract a best supporting sentence, Glaser and Schütze presented annotators with a single relative judgment task for each review. For each review, two sentences are presented to annotators. One sentence—the baseline sentence—is the review sentence that has been classified as sentiment-bearing with the highest-confidence by a sentence-level sentiment classifier, while the second sentence is the supporting sentence ranked highest by the supporting sentence weighting scheme. Annotators are then asked to choose the sentence that gives the more convincing reason (if neither sentence gave a convincing reason, annotators had the option of choosing *neither*). Approximately 1300 sentence pairs—one pair per review—were evaluated in this manner. Annotators rated the supporting sentence as the more convincing reason in 64.6% percent of cases.

8.2 Lexical features

8.2.1 Supporting argument lexicon-based features

Glaser and Schütze (2012) introduced a coarse-grained lexical feature scheme for *supporting sentence* classification: any nominal elements in a sentence previously classified as sentiment-bearing are considered good indicators that the sentence in question serves as a supporting sentence for the author's global sentiment. We would like to construct a more fine-grained set of lexical features motivated by the lexical characteristics of SAs described in the last chapter. In section 7.2.1, we noted that SAs can often be recognized by their use of explicit DMs such as *because*, *on the other hand*, and *if...then*, which serve to guide the reader's inference that the current segment functions as *Backing* or *Grounds* (using Toulmin's terminology) for a claim made elsewhere in the discourse context. Our approach to capturing the occurrence of DMs in an ICLE sentence involves compiling a lexicon of DMs identified by researchers as strongly associated with SAs. Our SA lexicon is composed of DMs extracted from two classes of metadiscourse markers compiled by Hyland (2005): *frame markers* and *transition markers*. Frame makers allow writers to sequence their arguments (*first of all, finally*), label argument stages (*in sum, to conclude*), or to announce argument goals (*in this part, intend to*), while *transition markers* include logical connectives such as *because* and *therefore*, along with additive DMs (*furthermore, in addition*) and contrast DMs (*on the other hand, nonetheless*). This gave us a small SA lexicon of 108 terms. SA lexicon lookup was implemented as a binary-valued feature: if a given sentence contained one of the items in the SA lexicon, the sentence received a value of True for this feature, otherwise, it received a value of False.

8.2.2 Stance and opinion word features

In section 3.6, we noted that stance is typically expressed using a variety of lexical forms that we collectively described as *evidential*. These evidential markers included deontic modals (*should, must*), predictive modals (*will, would*), and verbs of epistemic judgment (*indicate, think*). At the sentence level, evidential language is often used in ICLE essays to articulate a bare assertion, as in examples (19b) and (21b), repeated below as (89a,b), but they are also used to articulate SAs, as in (19a) and (21a), repeated below as (90a,b).

(89) a. Science technology and industrialization **clearly** dominates the world we live in. —*Science* prompt

b. I also **agree** that rehabilitation could be a nice alternative. —*Prisons* prompt

(90) a. Television is **certainly** people's Bible as far as information processing is concerned. —*Prisons* prompt

b. Another answer **might** be that the Norwegian politicians have not faced the fact that Cold-War is over. —*Armies* prompt

While the occurrence or frequency of occurrence of stancetaking lexis might be a good indication that a given sentence contains argumentative language of some kind, use of stancetekaing lexis does not necessarily tell us that a given sentence is a bare assertion or SA. Nevertheless, use of stancetaking lexis is ubiquitous in SAs, and a lexical feature set capturing the occurrence or frequency of occurrence of such lexis could serve as a potentially valuable feature in our *neutral-polar* SA classification experiments. In order to capture the occurrence of stancetaking language in a given essay sentence, we used the lexicon of stancetaking words constructed in section 5.2.5 to match any words in a given sentence to a stance word found in either the *for* or *against* sections of the stance lexicon. We then generalized each matched word by replacing all matches with the term `STANCE_WORD`.

As noted in Somasundaran and Wiebe (2010), stancetaking text contains a certain amount of opinion-bearing language, though the stance polarity and opinion polarity of a given text segment often do not overlap: one can use negative opinion-bearing language such as *old-fashioned* as part of a *for*-stanced argument, as in (33b), repeated below as (91).

(91) Initially I would say that the prison system is **old-fashioned**, especially when it comes to the rehabilitation of criminals. —*Prisons* prompt

We can exploit the lack of stance polarity/opinion polarity overlap evident in examples such as (91) as a classification feature. Given a prompt containing negative language, we would expect writers articulating a reason in support of a *for* stance toward that prompt to also use negative language and, vice versa, we would expect that reasons given in support of an *against* stance toward a prompt containing positive language would display positive language. We illustrate this point using three highly negative prompt texts. The majority of words in the *Prisons, Degrees*, and *Feminism* prompts are negative: the prison system is described as *outdated*, university degrees are described as *useless*, feminism has done *harm*, and so on. SAs used as part of a *for* stance toward any of these prompts will therefore tend to make use of negative opinion-bearing language (since the author will align herself with the negative opinion-bearing language of the prompt by using similarly negative language) while SAs used as part of an *against* stance will contain positive opinion-bearing language (since the author will disalign herself with the negative language of the prompt by using positive language). In examples (92-94), all three prompt texts are given, with all opinion words boldfaced, together with SAs found in both *for*- and *against*-stanced essays responding to these texts.

(92) PROMPT: All armies should consist entirely of professional soldiers: there is **no value** [*neg*] in a system of military service.

a. SA (*for*): Not only does military service **disrupt** [*neg*] daily life, especially when the person concerned doesn‘t have the makings of a soldier, but it may lead to a higher **death** [*neg*] rate.

b. SA (*against*): Some lower-educational people, and others for fulfilling their **dreams** [*pos*] of being a soldier, for them, it is a **proud** [*pos*], and they want to **devote** [*pos*] themselves to their country.

(93) PROMPT: Feminists have done more **harm** [*neg*] to the cause of women than **good** [*pos*].

a. SA (*for*): Enough is never enough and it might seem **pathetic** [*neg*] that some feminists still are so extremely **angry** [*neg*] -because women aren't represented 50% in the best paid jobs, quota above qualification.

b. SA (*against*): Women will be **respected** [*pos*] as well by **good** [*pos*] **achievements** [*pos*] what they received.

(94) PROMPT: Most university degrees are theoretical and do **not prepare** [*neg*] students for the real world. They are therefore of very **little value** [*neg*].

a. SA (*for*): Studying theoretical subjects takes a few years but seems to be **useless** [*neg*].

b. SA (*against*): To them a university degree is surely **valuable** [*pos*] for its very being theoretical.

How can we capture the opinion polarity of the SAs in (92-94) as a classification feature for our SA polarity classifier? We must first determine the dominant opinion polarity of each sentence. Rather than constructing a sentence-level opinion classifier from scratch, we make use of the coarse-grained approach to determining expression-level opinion polarity presented in Choi and Cardie (2009). We first encountered Choi and Cardie's approach in section 6.3, where it was used to replicate Somasundaran and Wiebe's method of determining sentence-level opinion polarity. Adopting the approach of Somasundaran and Wiebe, we extend Choi and Cardie's approach to whole sentences rather than just expressions.

A major component of Choi and Cardie's approach is the Vote and Flip algorithm, given in Figure 8.1, which was originally devised as a way of disambiguating the polarity of multi-word opinion expressions containing one or more negator words. Examples of such expressions include *not bad* and *never successful*—in the former case, *not* reverses the polarity of *bad* from negative to positive; in the latter case, the polarity of *successful* is reversed from positive to negative. As shown in Figure 8.1, the algorithm determines the positive, negative, or neutral polarity of the *i*th expression *e* by first setting the value of the variable *flipPolarity* to True or False depending on the number of negator words in an expression. If the number of negators is even, *flipPolarity* is set to False; otherwise, it is set to True (lines 7-10 of Figure 8.1). The value of *flipPolarity* is then used in

```
1: for each expression $e_i$ do
2:
3:     $numPositiveWords \leftarrow$ #positive words in $e_i$
4:     $numNeutralWords \leftarrow$#neutral words in $e_i$
5:     $numNegativeWords \leftarrow$ #negative words in $e_i$
6:     $numNegators \leftarrow$#negatiing words in $e_i$
7:     if ($numNegators$ % 2 = 0) then
8:         $flipPolarity \leftarrow False$
9:     else
10:        $flipPolarity \leftarrow True$
11:
12: if $(numPositiveWords > numNegativeWords)$ & $\neg flipPolarity$ then
13:     Polarity($e_i$) $\leftarrow positive$
14: else if $(numPositiveWords > numNegativeWords)$ & $flipPolarity$ then
15:     Polarity($e_i$) $\leftarrow negative$
16: else if $(numPositiveWords < numNegativeWords)$ & $\neg flipPolarity$ then
17:     Polarity($e_i$) $\leftarrow negative$
18: else if $(numPositiveWords < numNegativeWords)$ & $flipPolarity$ then
19:     Polarity($e_i$) $\leftarrow neutral$
20: else if $numNeutralWords > 0$ then
21:     Polarity($e_i$) $\leftarrow neutral$
22: else
23:     Polarity($e_i$) $\leftarrow default\_polarity$ (majority polarity in corpus)
```

Figure 8.1: Choi and Cardie's (2009) Vote and Flip algorithm

conjunction with opinion word counts (calculated using a lexicon of positive, negative, and neutral words), to determine the dominant opinion polarity of the expression. If none of the conditions in lines 12-21 of Figure 8.1 are met, the dominant polarity returned is the majority opinion polarity of the corpus. Since we do not know the majority opinion polarity of our corpus, our implementation of Vote and Flip returns a random polarity assignment for this final *else* condition. The Vote and Flip algorithm requires two knowledge sources, a lexicon of positive, negative, and neutral opinion-bearing words, and a list of negators. We used the MPQA subjectivity lexicon (Wilson and Wiebe, 2005) of opinion-bearing words and a single negator, *not* (along with the *n't* contraction). Once the dominant polarity of a sentence was determined using the Vote and Flip algorithm, all content words in the sentence were generalized to terms indicating that polarity (`POS`, `NEG`, or `NEUT`).

8.2.3 Capturing sentence-level topic information using Pointwise Mutual Information

Many Sentiment Analysis tasks require a reliable measure of association between two terms. These terms could be words, particular phrases, or structural features. A basic method of determining term association is Pointwise Mutual Information (PMI) (Jurafsky and Martin, 2009, pg. 26), which is formally defined as

$$PMI(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)},$$

where *p(x,y)* is the probability of terms *x* and *y* co-occurring (their joint distribution), while *p(x)* and *p(y)* are the independent distributions of *x* and *y*. The basic intuition motivating PMI is that two terms that co-occur more than is expected by chance are likely associated. The numerator indicates how often the terms do in fact co-occur and the denominator indicates how often they are expected to co-occur if we assume that each term occurs independently. The higher the resulting ratio, the stronger the association.

The PMI term association metric has been used in various ways in Sentiment Analysis. One common scenario involves using PMI to semi-automatically expand a lexicon of sentiment-bearing words by calculating the *n* words in a corpus that are most similar to particular seed words such as *good* or *bad* (Turney, 2002; Turney and Littman, 2003; Mullen and Collier, 2004). PMI has also been used in opinion classification tasks to discover opinion- or mood-denoting terms. As part of a blog mood classification task, Mishne (2005) used a search hit-count version of PMI to score the relationship between particular mood terms (*amused, depressed, ecstatic*, etc.) and blog-specific words. A search-hit count version of PMI was also used in the headline emotion classification work of Strapparava and Mihalcea (2008). In this work, the strength of emotion associated with content words found in headlines was scored by calculating the PMI between each content word and emotion-denoting terms such as *anger* and *surprise*. Other interesting work making use of the

PMI metric is the opinion target identification research described in Popescu and Etzioni (2005) and Jiang et al. (2011). In review text dealing with gadgets, opinions regarding the features associated with that gadget must somehow be linked to opinions regarding the product itself (e.g., opinions regarding features such as *interface* and *camera* must be linked to the target *iPhone*). Jiang et al. (2011) used a corpus of 20 million microblog posts to identify product features strongly associated with the targets of product reviews by calculating PMI scores for all nouns and noun phrases relative to product review targets.

Distribution-based word association metrics such as PMI can also provide a rough model of the topics discussed in a particular piece of text, which in turn can indicate the sentiment of that text. The ICLE sentences given in (95), extracted from essays written in agreement with the *Marx* prompt, both deal with the subject of indoctrination, using words commonly associated with this topic such as *controlled, power, deceived, passively, brainwashing*, and *ideas*.

(95) a. Just like an opium addict can be told anything, those who **controlled** society and had the **power**, could tell and **deceived** their subjects almost anything if they would only say it was in accordance with or even more, a duty according to their respective religions. —*Marx* prompt

b. Nowadays people accept **passively** this **brainwashing** and sets of **ideas** television directors are constraining to them. —*Marx* prompt

A PMI-based model of the topic (indoctrination) discussed in both these sentences would involve a set of PMI expansion words that overlap the topic space of the boldfaced words in (95).

To capture the association between topics and SA stance in ICLE essay sentences, we adopt the approach of Conrad et al. (2012) who experimented with sets of PMI-word expansions as part of a sentence-level stance classification task. Our first task was choosing a suitable corpus for our word distribution calculations. Since ICLE essays can deal with a variety of different topics, we needed a corpus of English text dealing with an unrestricted number of topic domains. We chose the 600 million word English section of *Wikicorpus* (Reese et al., 2010), a 750 million word,

ORIGINAL SENTENCE	NOUN (VERB)	PMI EXPANSION	
		EXPANSION	PMI SCORE
Adapting the old prison system would **cost** far too much.	cost	defrayed taxpayers estimation overruns minimization prohibitive	14.094 13.211 13.267 13.130 13.100 13.059
For instance when you read a book your **imagination** creates your own pictures and ideas.	imagination	foresight manifestations subconscious curiosity creativity reality	15.494 15.091 14.534 14.445 14.327 9.795
I can imagine that not many local authorities would want to take in a criminal who has committed five **murders**.	murders	overdosing suicides atrocious offenses misdemeanor abomination	15.088 14.999 14.180 14.470 14.348 14.064

Table 8.1: Examples of PMI-scored word expansions.

multilingual version of the online encyclopedia, *Wikipedia*.[3] After empirically determining that PMI scores based on five-word windows produced the best results, we calculated five-word window PMI scores between all nouns and verbs found in each ICLE sentence relative to all words in the English section of Wikicorpus. The topic word feature set associated with each sentence consists of the *n* highest-scoring PMI-based word expansions. In the experiments reported in section 9.3, we experiment with 5 and 15 PMI word expansion sets. Examples of high-scoring PMI-scored words extracted in this manner are given in Table 8.1, with the original sentence from which the noun (verb) was extracted given in the first column, the extracted noun (verb) given in the second column, and a selection of high-scoring expansion words and their associated PMI scores given in the third and fourth columns.

[3] http://en.wikipedia.org/wiki/Main_Page

8.3 Discourse and rhetorical structure features

While discourse structure has historically played a minor role in document- and sentence-level sentiment classification research, it has played a key role in several Sentiment Analysis sub-tasks. These include sentiment polarity disambiguation (Polanyi and Zaenen, 2006; Zhou et al., 2011), weighting of sentiment-bearing words and sentences by their occurrence in discourse role-specific text spans (Voll and Taboada, 2007; Taboada et al., 2008; Heerschop et al., 2011), and sentence-level classification of argument types (Conrad et al., 2012).

Polanyi and Zaenen (2006) suggested that discourse structure could help to disambiguate the sentiment polarity of sentences such as (96), which exhibits an intra-sentential "contextual valence shift" of *contra-expectation* signaled by the use of *although*.

(96) Although Boris is **brilliant** at math, he is a **horrible** teacher.

Use of *although* together with the negative sentiment-bearing *horrible* in the second clause of (96) negates the positive sentiment (signaled by *brilliant*) in the first clause. Such sentences are by no means rare in sentiment-bearing text. Zhou et al. (2011) discovered that a full 43% of the sentences in the NTCIR MOAT (Multilingual Opinion Analysis Task) Chinese corpus are polarity-ambiguous. Zhou et al. incorporated sentence-level discourse relationship information into the sentiment-tagged MOAT corpus, achieving a significant boost in accuracy over a standard sentence-level sentiment classification baseline. Using a small set of discourse cue-based patterns bootstrapped from a subset of the MOAT corpus, Zhou et al. constructed a coarse-grained discourse parser based on the Rhetorical Structure Theory (RST) scheme of Mann and Thompson (1988) with the full set of RST relations reduced to a handful of basic relations (*Contrast, Condition, Continuation, Cause*, and *Purpose*). Information from a sentence's discourse tags was incorporated into the final feature set by writing a set of constraints for each discourse tag. A sentence such as (96), for example, which would be tagged as *Contrast*, contains two segments of opposite polarity, while a sentence tagged as *Continuation* contains two segments with the same polarity. With these constraints added to the baseline classification model, the final classification model presented in Zhou et al. achieved an F-score of of 81.0% relative to an F-score of 76.4% achieved using only opinion polarity features.

Sentiment Analysis research involving the use of discourse structure information to selectively weight particular sentences or words in a document is motivated by the intuition that certain sections of a sentiment-bearing document, such as a review or editorial, are more indicative of that document's global sentiment than others. Taboada and Grieve (2004) assigned weights to sentiment-bearing words based on their location in a review text, with words located at the beginning of the text receiving the lowest weight, since introductory sections of reviews tend to report background information rather than direct evaluations of the company or product. The sentiment-bearing words found toward the end of the review received the highest weight, since reviews tend to end with summary or encapsulating statements of the author's opinion. Taboada and Grieve (2004) used this approach as part of a document-level sentiment classification task and reported an overall accuracy of 65% relative to a 51% baseline.

Other Sentiment Analysis research involving the selective weighting of words or sentences based on their position is found in the study of Taboada et al. (2008). Taboada et al. used the SPADE (Soricut and Marcu, 2003) sentence-level discourse parser to assign RST relations to sentences in a corpus of *epinions.com* reviews. RST assigns *nucleus/satellite* relations to text spans, with the nucleus span serving as the central span and the satellite serving as the supporting span. Each nucleus/satellite relation is marked as a discourse relation (*Concession, Condition, Evidence*, etc.). Taboada et al. selectively weighted sentiment-bearing words based on their occurrence in a nucleus or satellite, with terms that occurred in nuclear text spans receiving a higher weight. After weighting terms in this manner Taboada et al. report an accuracy of 80.0% relative to an accuracy of 72.0% achieved without selective weighting.

Finally, Heerschop et al. (2011) used sentence-level RST discourse relations to selectively weight sentiment-bearing words based on their occurrence in particular RST-tagged spans. The approach of Heerschop et al. is more fine-grained than that of Taboada et al. in that weights were assigned based on specific RST relations types (*Concession, Condition,* etc.) rather than simple nuclei/satellite relations. Their highest reported accuracy is 72.0%, which significantly improves upon the 68.8% accuracy achieved by a baseline system that ignores discourse structure information.

8.3.1 Constructing discourse relation features using the Penn Discourse Treebank

Our approach to incorporating discourse structure information into our feature sets is partly inspired by Conrad et al. (2012), who used discourse relations based on the Penn Discourse Treebank (PDTB) of Prasad et al. (2008) as part of a sentence-level classification task. Conrad et al. collected 84 examples of blog articles and editorials dealing with the healthcare reform debate. These documents were then segmented into 2678 sentences for use in two distinct classification tasks: the classification of a sentence as displaying "arguing subjectivity" (i.e., displaying a stance) and the classification of any stanced sentences according to the argument type articulated. These argument types were limited to a single, restricted domain—stancetaking text dealing with proposed healthcare reform—and include such topic-specific arguing types as *improves healthcare access* and *expands government*. Conrad et al. used the recently released PDTB parser of Lin et al. (2010) to incorporate PDTB relation information into their feature set. Additional features included binary-valued features indicating the presence/absence of sentiment-bearing language and sets of semantically similar words. Surprisingly, the PDTB feature set did not significantly improve accuracy above a 61.0% bag of words baseline for the recognition of arguing subjectivity task. Instead, the best-performing feature combination for this task included sets of 10 semantically similar words extracted from a corpus of Wikipedia articles and the binary-valued sentiment-bearing language feature. Despite this negative result, we elected to experiment with a

PDTB-based feature set, reasoning that the very different text variety used in current experiments contained predictable structural characteristics that could be captured by the PDTB parser.

The publicly available parser used to construct our feature set was trained on the PDTB and is described in Lin et al. (2010). Lin et al. report an F score of 86.7% for the explicit relation classification component of their PDTB parser and an F-score of 39.6% for the far more difficult task of automatically classifying implicit discourse relations. We can view their parser, then, as a fairly reliable tool for parsing the ICLE essays used for the supporting argument classification work reported here.

As described in Prasad et al. (2008), the PDTB is a compete set of discourse relations annotated over the 1 million word Wall Street Journal Corpus. The discourse relations themselves are theory-neutral binary relations, or "senses," connecting two arguments: *Arg2*, which is the text segment to which the connective is syntactically bound, and an additional argument, *Arg1*. The PDTB maintains a top-level division between *explicit* discourse relations, which are realized by many of the DMs discussed in section 7.2, such as *because* and *although*, and *implicit* discourse relations, which lack explicit DMs and rely instead on the contiguity of *Arg1* and *Arg2* to signal a discourse relationship. Since arguments can be hierarchically embedded within other arguments (i.e., an *Arg1/Arg2* pair can itself serve as a member of an argument pair) both within and across sentences, the resulting discourse structure can often be complex. To illustrate, consider the two-sentence excerpt given in (97).

(97) This, of course, makes us rational and does not leave us much time nor place for dreaming in the everyday life. However, imagination exists, although it is suppressed.

This excerpt contains two discourse relations, one relating both sentences in a relation of *contrast* (signaled by the DM *however*) and the other relating two segments of the second sentence in a relation of *concession* (signaled by *although*). Figure 8.2 gives the PDTB parse of (97), showing both inter- and intra-sentential relations.

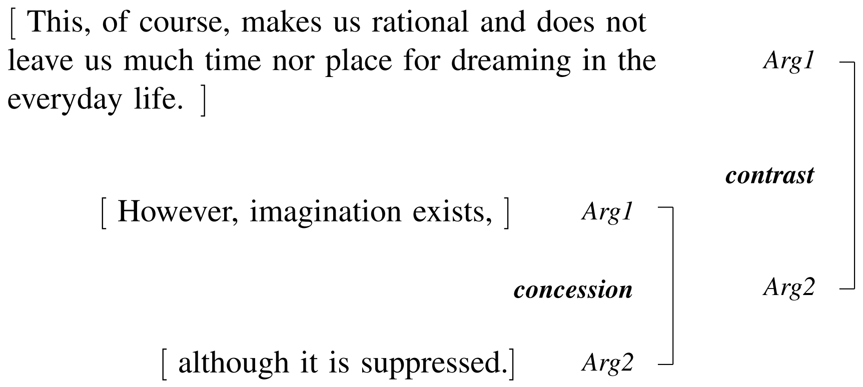

[This, of course, makes us rational and does not leave us much time nor place for dreaming in the everyday life.]
Arg1
contrast
[However, imagination exists,]
Arg1
concession
Arg2
[although it is suppressed.]
Arg2

Figure 8.2: PDTB parse of the two-sentence essay excerpt given in (97).

The PDTB parser of Lin et al. reduces the 100+ discourse relation types captured in the PDTB to a more manageable size by subsuming many relation sub-types under parent relation types such as *Cause, Condition, Contrast, Concession, Restatement*, etc. Thus, discourse relations of *specification, equivalence,* and *generalization* are subsumed under their parent relation type *Restatement*, relations of *reason* and *result* are subsumed under *Cause*, and so on. The resulting tag set has 16 explicit relation tags, and a single non-explicit relation tag that can be combined with 13 of the 16 explicit relation tags, resulting in a total of 29 distinct relation tags. After parsing the entire ICLE corpus, we found that only 20 of these tags were attested in the parsed essays. Our complete set of PDTB-based features, then, consists of 20 binary-valued PDTB tags. For each tag, a sentence receives a value of True if it was assigned the tag by the parser or False if it was not assigned the tag by the parser.

Our hypothesis is that both intra- and inter-sentential PDTB parse tags can help to identify several key discourse structure characteristics of supporting arguments in our *neutral-polar* classification experiments. Additionally, we hypothesize that such information can boost the accuracy of the *neutral-polar* supporting argument classification model described in section 9.2. One might ask if a PDTB-based feature set is redundant since our SA lexicon-based feature set is already designed to capture important discourse relations, such as *Contrast* and *Concession*, by noting the occurrence of DMs such as *because* and *although*. In answering this question, we first consider the PDTB-parsed ICLE sentences given below in (98-100). In (98) and (99), *Cause* relations are displayed that are already captured by the SA lexicon-based feature described in section 8.2.1 (the causal DMs *because* and *therefore* are in our SA lexicon). The sentence pair in (100), which involves an intra- and inter-sentential discourse relationship, shows that subtle, discourse structure-based characteristics of SAs often cannot be recognized via obvious lexical cues such as *when* and *because*.

(98) ***Cause***

According to me, the modern technologically dominated world has made $[_{\substack{\text{Explicit Arg1}\\ \text{Cause}}}$ us feel more isolated and more desperate] **because**$_{\textbf{Cause}}$ $[_{\substack{\text{Explicit Arg2}\\ \text{Cause}}}$ we have lost our vital connection with our soul and our contact with the people that surround us.]

(99) ***Conjunction, Cause***

$[_{\substack{\text{Explicit Arg1}\\ \text{Cause}}}$ $[_{\substack{\text{Explicit Arg1}\\ \text{Conjunction}}}$ Nowadays we can see and evaluate the innumerable triumphs of the human mind and exactly that inclines us to think that its complexity and inventiveness are infinite]] **and**$_{\textbf{Conjunction}}$ $[_{\substack{\text{Explicit Arg2}\\ \text{Conjunction}}}$ **therefore**$_{\textbf{Cause}}$ $[_{\substack{\text{Explicit Arg2}\\ \text{Cause}}}$ humans can move on only towards perfection in science, technology and art]]

(100) ***Cause, Instantiation, Synchrony***

a. $[_{\substack{\text{Non-explicit Arg1}\\ \text{Instantiation}}}$ $[_{\substack{\text{Explicit Arg1}\\ \text{Cause}}}$ That is so] **because**$_{\textbf{Cause}}$ $[_{\text{Explicit Arg2Cause}}$ things in theory are quite different from things in practice.]]

b. $[_{\substack{\text{Non-explicit Arg 2}\\ \text{Instantiation}}}$ **When**$_{\textbf{Synchrony}}$ $[_{\substack{\text{Explicit Arg2}\\ \text{Synchrony}}}$ one receives their education] $[_{\substack{\text{Explicit Arg1}\\ \text{Synchrony}}}$ he or she learns many things that will be of no practical use to him or her in their respective jobs.]]

(100b) is an SA provided as *Backing* for an argument endorsing the Degrees prompt's claim that "most university degrees are theoretical and do not prepare students for the real world." The PDTB parser identifies (100b) as an (implicit) *Instantiation* of the argument made in (100a). In the PDTB scheme, an instantiating text segment expands upon another text segment by providing additional details or reasons. Unlike (98) and (99), we do not find explicit use of DMs used to identify (100b) as an SA. In addition, the PDTB parser assigns an intra-sentential tag of *Synchrony* to this sentence, cued by the use of *when*. In the PDTB scheme, two text segments are synchronous when the situation described in the second segment is contingent upon the first, either logically or temporally. In this case, the situation described in the sub-clause (the learning of many things that will be of no practical use) is contingent upon the situation described in the main clause (going to university). Synchrony is a subtle intra-sentential discourse strategy commonly used to articulate SAs; however, this strategy cannot be captured as a feature using lexicon-based methods—DMs indicating synchrony, such as *when* and *while*, are far too polysemous to be included as reliable indicators of SAs in our SA lexicon.

8.3.2 Rhetorical stage feature

In section 7.3.2, we described Hyland's move analysis of student argumentative essays. As given in Table 7.3, Hyland's rhetorical organization scheme divides argumentative essays into three top-level stages, Thesis, Argument, and Conclusion. In trying to capture Hyland's taxonomy as a classification feature, we restrict our attention to these top-level categories. The use of paragraph- or section-specific sentence location as a learning feature has a lengthy history in NLP, particularly in areas such as text summarization and *argumentative zoning* (i.e., the automated identification of argumentative stages in scientific research articles). In early work involving the automated indexing of technical literature, Baxendale (1958) noted that sentences located at the beginning and end of paragraphs in a technical article contain language that is strongly indicative of the overall topic of that article. This insight was later implemented in the text summarization work of Brandow et al. (1995), whose topic word metric gave prominence to words contained in the leading sentences of paragraphs, and Hovy and Lin (1998), who noted that the first sentence of the second paragraph of news articles is strongly indicative of an article's topic. In work dealing with the argumentative zoning of research articles, Teufel's (1999) feature set included the relative position of a sentence in particular sections of research papers.

We determine whether a given sentence is located in a Thesis, Argument, or Conclusion section of an essay in the following manner. First, all paragraph boundaries in an essay are located using the Montylingua toolkit's (Liu, 2004) paragraph segmenter. If the essay contains three or more paragraphs, then the first and last are considered the Thesis and Conclusion, respectively, while the rest of the essay is considered the Argument. If the essay does not contain any paragraph breaks, then the entire essay is split into three sections, with each section containing the same number of sentences. Each of these three sections is then considered a Thesis, Argument, or Conclusion section.

8.4 Summary

In this chapter, we described a set of classification features motivated by the lexical, discourse structure, and rhetorical structure characteristics of SAs described in chapter 7. Two distinct sets of features were described: a set of features designed for the *neutral-polar* classification experiments described in section 9.2, and a set of features designed for the polarity classification experiments described in section 9.3. At the lexical level, a sentence can often be recognized as an SA by noting the use of explicit DMs, such as *because* and *although*, as described in Fraser (2006). To capture this aspect of SAs, we compiled a lexicon of 108 DMs from the list of metadiscourse markers collected in Hyland (2005). This lexicon is used to match any DMs found in the sentence and the SA lexicon feature itself is binary-valued: it is set to True if a match is found; otherwise, it remains False.

We then described lexical features designed to capture the occurence of stancetaking and opinion-bearing language in SAs. As described in 8.2.2, the occurrence of stancetaking language in a sentence is captured by matching any stancetaking words in the sentence to words in the stance lexicon described in section 5.2.5 and then generalizing those matched words to STANCE_WORD. The occurrence of opinion-bearing language in SAs is captured in our polarity classification feature set by first determining the dominant polarity of the sentence (using Choi and Cardie's Vote and Flip algorithm) and then generalizing all content words in the sentence to the tag associated with that polarity—POS, NEG, or NEUT.

In section 8.2.3, we described the construction of sets of PMI-scored words which are designed to capture the relationship between the topic of an SA and the polarity of that SA. 5-word windows were used to calculate the similarity between all nouns and verbs in the sentence and words in a 600 million word corpus of English text, *Wikicorpus* (Reese et al., 2010). For each noun and verb in each sentence, the 5 and 15 highest-scoring words are represented as two separate features. An additional feature included in the polarity classification feature scheme is the stance of the essay from which the sentence was extracted, as determined using the essay-level stance annotations described in chapter 4.

Discourse structure features designed for *neutral-polar* classification included a set of features based on the discourse relationship senses of the PDTB. The PDTB parser of Lin et al. (2010) was used to discourse parse all sentences in the ICLE corpus and each sentence's discourse sense tags were represented as a set of 20 binary-valued features. The final *neutral-polar* classification feature was described in section 8.3.2, and is designed to incorporate information regarding the global rhetorical structure of the essay into the *neutral-polar* classification feature set. The rhetorical structure feature is based on the essay organization framework of Hyland (1990), who represented the student argumentative essay as a series of rhetorical stages. Hyland's three top-level stages of Thesis, Argument, and Conclusion are captured as a nominal-valued feature by noting the position of a given sentence in one of these stages in the original essay from which that sentence was extracted.

Chapter 9

Supporting argument classification experiments

In this chapter, we supplement the essay-level stance classification model described in chapters 5 and 6 with two sentence-level classification models. The first, *neutral-polar* model identifies a given sentence as an SA. The second model classifies the polarity of an SA as either *for* or *against*. Although these models are not presented in a two-stage manner, they can be used in this way in applications. In our final chapter, we present one such application: a single sentence summarization model that first identifies any SAs in an ICLE essay (using the first, *neutral-polar* model) and then classifies all sentences identified as SAs as holding either a *for* or *against* stance (using the second, polarity classification model). The SA classified with the highest confidence is considered the most representative SA used in that essay and serves as a single-sentence summary of the essay's arguments backing a *for* or *against* stance toward the essay prompt.

9.1 Supporting argument annotation

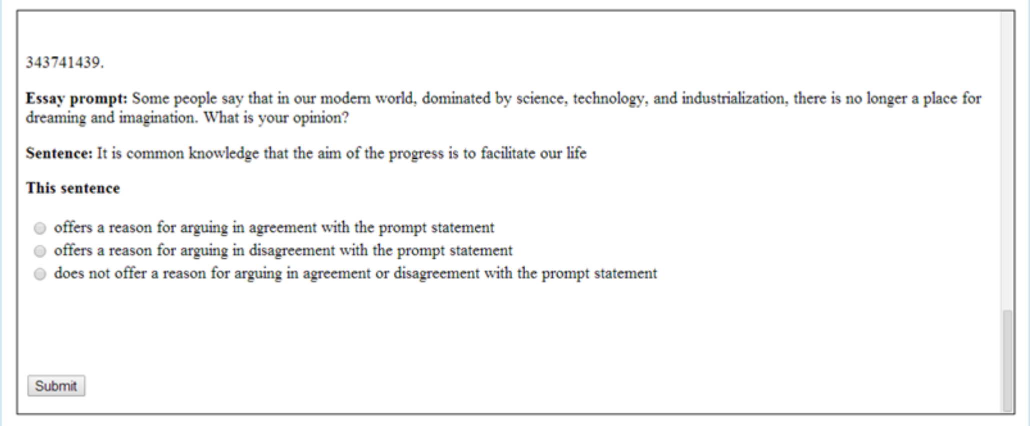
9.1.1 Background

The most comprehensive example of sentence- and phrase-level stance annotation is the *arguing subjectivity* annotation work described in Wilson (2008). This stance annotation work was one of several sentiment annotation tasks completed for the MPQA Opinion Corpus of Wilson and Wiebe (2005). The MPQA Opinion Corpus is comprised of news text from various sources annotated for *private states*. In the context of the MPQA annotation scheme, a private state is a general term for beliefs, thoughts, feelings, emotions, evaluations, and judgments. Annotators were asked to mark spans of text that in their judgment expressed various private states such as attitude (positive, negative, other, or none) and intensity (low, medium, high, or extreme). A recent addition to the MPQA Opinion Corpus is the inclusion of positive and negative-arguing annotation schemes to capture spans of text denoting private states in which the author or quoted speaker is "expressing a belief about what is true or should be true in his or her view of the world." (Wilson, 2008, pg. 117). This description of arguing subjectivity is compatible with the definition of stancetaking language that we have maintained throughout this study: stancetaking language allows speakers/writers to endorse or disclaim the truth of likelihood of a proposition. Wilson (2008) reports the results of an agreement study involving the identification of spans of text containing fine-grained attitudes such as *agreement, arguing*, and *speculation*. Overall agreement regarding attitude types was quite high: observed agreement and Cohen's *k* were .86 and .78, respectively. In the case of arguing subjectivity, annotators agreed in 145 of 156 cases that given text spans contained language of this attitude type.

Other annotation work involving sentence-level stance annotation is reported in Conrad et al. (2012). This annotation work involved identifying text spans as displaying one of 18 *for* or *against* topic-specific argument types dealing with the current debate regarding healthcare reform. Each of these argument types represents a conceptual strand typical of arguments made *for* or *against* the proposed healthcare initiative. For example, arguments asserting that unemployment will rise as a result of the proposed reforms or that the reforms will curtail business expansion were grouped under the *against* arguing type *hurts_economy*, while arguments that senior citizens and the uninsured would benefit from the proposed reforms are grouped under the *for* argument type *improves_healthcare_access*. Two annotators tagged 384 sentences according to this scheme. Two levels of agreement were evaluated: agreement regarding the existence of stancetaking language in a text span and agreement regarding the arguing type of the language contained in those stancetaking text spans. Conrad et al. report an F-score of .68 for the former level of agreement and a Cohen's *k* of .68 for the latter level of agreement.

9.1.2 Corpus processing and annotation steps

When selecting sentences for the SA stance annotation task described in this section, we ensured that the prompt distribution of the essays from which the sentences were extracted matched the general prompt distribution of the ICLE corpus given in Table (4.1) (with *for* and *against* essays evenly divided for each prompt-specific subset of essays). This means that, of the 239 essays used in the SA annotation task, 28% (=68) are essays responding to the *Science* prompt (since 28% of the ICLE corpus consists of *Science* essays), 11% (=26) are essays responding to the *Feminism* prompt (since 11% of the ICLE corpus consists of *Feminism* essays), and so on. All essays were segmented into individual sentences using the MontyLingua toolkit (Liu, 2004). This resulted in an annotation set of 8176 sentences. All sentences were posted to AMT and three unique annotators were assigned to each sentence. Annotators were provided with 10 sentences to annotate per screen, with sentences randomly selected from the original set of 8176. For each sentence, the prompt associated with the essay from which that sentence was extracted was provided. As described in the annotation protocol given in Appendix B, annotators were instructed to read the prompt in its entirety, followed by the accompanying sentence, and then to decide if the sentence offered a reason for arguing *for* the prompt statement, offered a reason for arguing *against* the



343741439.

Essay prompt: Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?

Sentence: It is common knowledge that the aim of the progress is to facilitate our life

This sentence

○ offers a reason for arguing in agreement with the prompt statement
○ offers a reason for arguing in disagreement with the prompt statement
○ does not offer a reason for arguing in agreement or disagreement with the prompt statement

Figure 9.1: Screenshot of the AMT interface for the SA annotation task.

		Amazon Mechanical Turk			
		FOR	AGAINST	NEITHER	TOTAL
GOLD	FOR	**1058**	38	373	1469
	AGAINST	66	**969**	253	1288
	NEITHER	250	229	**4940**	5419
	TOTAL	1374	1236	5566	**8176**

Table 9.1: Contingency table for the supporting argument stance annotation task.

prompt statement, or did not offer a reason for arguing either *for* or *against* the prompt statement. As with the essay-level stance annotation task described in chapter 4, all gold-standard annotations were completed by the author and annotation tags for the AMT-tagged corpus were determined using majority voting and random tie-breaking. A screenshot of the AMT interface used for this task is given in Figure 9.1.

Observed agreement and Cohen's *k* for this annotation task was .85 and .70, respectively. Examining the results of this task, given as a contingency table in Table 9.1, we can see that the comparatively high observed agreement score is largely due to the massive number of *neutral* sentences included in the annotation set—fully 66% (=5419 sentences) of the gold-standard annotation set consists of *neutral* sentences, which annotators found easy to identify.

9.1.3 Error analysis

Examining Table 9.1, we find that annotators generally had good intuitions regarding the *neutral* or *polar* status of sentences: of 2757 polar sentences, 77% were tagged as *polar* by annotators. There is also high observed agreement for *neutral* sentence annotation: over 90% of *neither* sentences were tagged *neither* by annotators. Consistent sources of disagreement regarding the *neutral* or *polar* status of sentences involved sentences such as (101). In (101a), the author's negative reaction to the prompt statement is interpreted as a reason for arguing *against* that statement (a version of the *argumentum ad passiones*, or appeal to emotion fallacy). In (101b), a deontic assertion is interpreted as a reason for arguing *against* the prompt.

(101) a. GOLD=*neither*, AMT=*against*: If we look at the historical background for the feminist movement, it is difficult to take a statement as Feminists have done more harm to the cause of women than good seriously. . —*Feminism* prompt

b. GOLD=*neither*, AMT=*against:* We should follow our heart dictates and let our minds free for a while for reflection. —*Science* prompt

Cases of inter-polarity disagreement often involved varying intuitions regarding the correct *Warrant* (using Toulmin's terminolgy) to be inferred given examples of *Backing* statements such as (102a). The *when...they* structure of (102a) is an example of *synchrony*, using the terminology of the PDTB; cf. section 8.3.1. The situation presented in the *they*-sub-clause is presented as a logical consequence of the situation presented in the *when*-main-clause. (102a) serves as an argument in support of a *for* stance toward the main proposition in the *Money* prompt, *Money is the root of all evil*, since the author asserts that the pursuit of wealth has negative moral consequences. Correctly interpreting this argument as a *for*-stanced reason involves inferring the *Warrant* that *Lack of respect, peace, and love is associated with evil*. Similarly, in (102), there is disagreement regarding the implied consequences of the author's claim that television can help to keep politicians accountable. One consequence of this claim is that television can empower its viewers—since this is the opposite of its role as a tool of indoctrination, the gold-tag for this sentence is *against*, though

AMT annotators had difficulty making this inference and tagged this SA as *for*.

(102) a. GOLD=*for*,AMT=*against*: However, when they try to gain money, they become to forget some good behaviors or feelings such as respect, peace or love. —*Money* prompt

b. GOLD=*against*, AMT=*for*: Here one can see that television may serve as a powerful tool to correct politicians. —*Marx* prompt

9.1.4 Discussion

As with essay-level stance annotation, the high levels of inter-annotator agreement for the SA classification task suggest that even non-expert annotators have very clear intuitions regarding both the existence of argumentative language in a given sentence and the stance polarity of that sentence. A similar result is reported in Wilson (2008). Cases of neutral-polar disagreement often involved appeals to emotion and deontic assertions, both of which were incorrectly tagged as polar by annotators. These issues could be dealt with in future expert annotation work by training annotators to recognize such sentences, which in turn involves training annotators to recognize patterns of informally valid and invalid reasoning. While the computational modeling of formal argumentative and deductive logical schemes is as old as Artificial Intelligence itself (Russell and Norvig, 2010, pg. 4), there exist only a handful of approaches in NLP that deal with informal argumentative schemes (Palau and Moens, 2009; Feng and Hirst, 2011). A professionally annotated corpus of stance-annotated sentences, then, could be of some benefit to researchers in stance classification and various other fields involved in the modeling of informal, or “natural language arguments” (Reed and Grasso, 2001).

FEATURE	DESCRIPTION
PDTB-based features	20 binary-valued PDTB senses. For each sense, if the sentence received the PDTB parser tag associated with that sense, the value of the sense is set to True; otherwise, it remains False.
SA-lexicon based feature	Binary-valued feature. If the sentence contains a DM contained in the SA lexicon, the value of the feature is set to True; otherwise, it remains False.
Hyland stages-based feature	Nominal-valued feature indicating the sentence's position in the Thesis, Argument, or Conclusion section of the essay.
Stance-generalized unigrams	All sentences are transformed into vectors of words. Each word matched to an item in the stance lexicon is generalized to `STANCE_WORD`. Each word is represented as a separate feature that is either frequency-valued (in the case of Naive Bayes and the SVM classifier) or binary-valued (in the case of Multivariate Bernouli Naive Bayes).

Table 9.2: Features used in the neutral-polar classification experiments.

9.2 Neutral-polar classification: Experimental setup and results

Of the 8176 sentences annotated for SA stance, 1796 were used for the *neutral-polar* classification experiments reported in this section. Of these 1796, 917 (=51%) were neutral and 879 (=49%) were *polar* (439 *for* sentences and 440 *against* sentences). The *neutral-polar* classification features given in Table 9.2 were used in all experiments.

We used the machine learning algorithms provided in the Weka machine learning toolkit (Hall et al., 2009) for all experiments. While experimenting with our development set, we found that Weka's implementation of NB achieved better results than NBM and so we use the former learning algorithm for both *neutral-polar* and *polar* classification experiments. To ensure that the performance of our feature set is not an artifact of a particular classifier, we compare the performance of NB with NBMB and an SVM classifier. During experimentation with our development set, we found that a linear, rather than RBF kernel, performed best for the SVM classifier. We experimented with different combinations of the features given in Table 9.2 by cumulatively adding features atop a majority-tag model. All sentences were normalized using the normalization function described in section 6.3. Our baseline mode for each classifier is a bag-of-words model, i.e., an unordered vector of frequency-valued (in the case of the NB and SVM classifiers) or binary-valued (in the case of the NBMB classifier) unigrams. Table 9.3 provides the complete set of neutral-polar classification experiment results. The significance of any increase in accuracy resulting from cumulatively adding additional features to the model was measured using McNemar's χ^2 test at level $p < .05$.

		Polar			**Neutral**		
	Acc.	Prec.	Rec.	F	Prec.	Rec.	F
Naive Bayes							
Baseline: Bag of words	66.4	65.8	67.5	66.6	67.1	65.4	66.3
None (majority tag)	51.0	0	0	0	51.1	1	67.6
PDTB	56.6*	60.8	32.1	42.0	55.1	80.1	65.3
PDTB + SA	57.4	59.9	39.5	47.6	56.3	74.7	64.2
PDTB + SA + Hyland	57.4	59.9	39.5	47.6	56.3	74.7	64.2
PDTB + SA + Hyland + Stance gen. words	**70.1***	67.0	77.1	71.7	74.3	63.5	68.5
Multivariate Bernouli Naive Bayes							
Baseline: Bag of words	69.8	73.3	61.8	67.1	67.4	77.8	72.2
None (majority tag)	51.0	0	0	0	51.1	1	67.6
PDTB	55.8*	59.2	31.4	41.0	54.7	79.3	64.7
PDTB + SA	55.8	59.2	31.4	41.0	54.7	79.3	64.7
PDTB + SA + Hyland	56.5	58.5	38.5	46.4	55.6	73.8	63.4
PDTB + SA + Hyland + Stance gen. words	**74.3***	77.4	67.3	72.0	72.2	81.1	76.4
SVM (linear kernel)							
Baseline: Bag of words	67.7	70.1	61.0	65.3	65.9	74.3	69.9
None (majority tag)	51.0	0	0	0	51.1	1	67.6
PDTB	52.3*	67.1	5.3	9.9	51.8	97.5	67.7
PDTB + SA	52.9*	68.1	7.3	13.2	52.1	96.7	67.7
PDTB + SA + Hyland	52.9	68.1	7.3	13.2	52.1	96.7	67.7
PDTB + SA + Hyland + Stance gen. words	**71.6***	77.3	59.4	67.2	68.2	83.3	75.0

Table 9.3: *Neutral-polar* supporting argument classification experiment results. For each classifier, the highest accuracy is boldfaced. The highest accuracy across all three classifiers is boldfaced and underlined. * indicates a statistically significant increase in accuracy at level $p < .05$ (using McNemar's test) relative to the feature set immediately above.

9.2.1 Discussion

Each classifier in Table 9.3 achieved its highest accuracy using the complete set of features, with the NBMB classifier scoring highest (74.3%), followed by the SVM (71.6%) and NB (70.1%)

classifiers. These three classification models achieved accuracies that were significantly higher than their respective bag-of-words baselines at level $p < .01$, as measured using McNemar's test. For all three classifiers, the PDTB features contributed to significant gains in accuracy at level $p < .05$, above a majority-tag model. For the NB and NBMB models, PDTB sense-based features contributed to a 10% and 8% increase in accuracy, respectively, over the majority-tag model. This contrasts with the negative result reported in Conrad et al. for this feature set. Our result seems to confirm the hypothesis that PDTB-based senses can serve as valuable sentence-level features for stancetaking text if the text itself is drawn from a domain, such as student argumentative essays, that involves formulaically structured argument patterns.

Manual examination of the NBMB model revealed that PDTB senses were somewhat more useful in identifying a sentence as stanced (*polar*) than as not stanced (*neutral*). In particular, Explicit *Cause* senses served as helpful discriminators of polar sentences. In the PDTB scheme, the *Explicit Cause* sense captures those constructions in which a DM such as *since* or *because* indicates that the situation in *Arg2* has caused (temporally or logically) the situation in *Arg1*. *Cause* sub-types such as *reason* are paradigmatically associated with SAs. In (103a),the explicit causal DM *because* introduces an SA backing an *against* stance toward the *Degrees* prompt while in (103b) *because* is used as part of an SA backing a *for* stance toward the *Science* prompt.

(103) a. **PDTB SENSE=EXPLICIT CAUSE/ TAG=POLAR:** Because the more you learn the better salaries you will get

b. **PDTB SENSE=EXPLICIT CAUSE/ TAG=POLAR:** The way I see it there is no longer a place for our dreams and imaginations to come true, because of science technology and industrialisation.

Turning to the results of the SVM classifier, we find in Table 9.4 that the highest-weighted PDTB senses learned by the SVM classifier (relative to the polar class) included the *Restatement, Instantiation*, and *Contrast* senses. As mentioned in section 8.3.1, the *Restatement* sense captures those constructions in which the situations described in *Arg1* and *Arg2* both hold at the same

Polar TERM WEIGHTS

PDTB SENSE	WEIGHT
Non-explicit Arg2 Restatement	0.616
Explicit Alternative	0.5406
NonExplicit Arg2 Instantiation	0.3285
Non-explicit Arg2 Contrast	0.3106
Explicit Restatement	0.271
Explicit Asynchronous	0.1471
Explicit Contrast	0.1372
Non-explicit Arg2 Cause	0.127
Explicit Instantiation	0.023

Neutral TERM WEIGHTS

PDTB SENSE	WEIGHT
Explicit Concession	-0.8312
Non-explicit Arg2 No Relation	-0.8182
Explicit Cause	-0.3634
Non-explicit Arg2 Entity Relation	-0.2091
Expliic Synchrony	-0.1848
Explicit Conjunction	-0.1779
Explicit Condition	-0.1649
Non-explicit Arg2 Conjunction	-0.0075
Explcit t List	-0.02
Non-explicit Arg2 Asynchronous	0

Table 9.4: PDTB sense weights learned by the SVM neutral-polar classifier

time. The implicit *Restatement* sub-type *equivalence* is the counterpart of arguments introduced by the DM *in other words* as shown in (104a), where the writer expands upon the scenario given in the italicized *Arg1* by introducing an equivalent scenario in *Arg2*. Similarly, in the implicit *Instantiation* construction given in (104b), the scenario in *Arg2* expands upon or adds further detail to the claim made in the italicized *Arg1*. Finally, the explicit *Alternative* relation sense in (104c) is triggered by the use of *or* which writers often use in SAs to present positive or negative scenarios that back a stance toward the prompt statement.

(104) a. *Nowadays the best friend of man is no more the dog but the computer.*
PDTB SENSE=NON-EXPLICIT ARG 2 RESTATEMENT/TAG=POLAR [EQUIVALENCE/ “In other words”]: In that hostile world that has killed your pure dreams,

the computer can easily become your become your preferred company.

b. *It should be mentioned that political situation is very important.*
PDTB SENSE=NON-EXPLICIT ARG 2 INSTANTIATION/TAG=POLAR: In Russia, the professional army can become dangerous, used by any ambitious person who has a great desire to get power.

c. **PDTB SENSE=EXPLICIT ALTERNATIVE/ TAG=POLAR:** Is it really possible to rehabilitate a man who has blood of a murderer **or** who is mentally ill?

Surprisingly, the SA lexicon-based feature and the Hyland stages-based feature performed poorly across all three classifiers. The SA lexicon feature contributed significantly to a slight increase in accuracy in the SVM but failed to contribute to significant accuracy increases in either the NB and NBMB classifiers. The Hyland feature failed to contribute to increases in accuracy across all three classifiers. The poor performance of the SA lexicon feature can be attributed to the massive polysemy of explicit DMs (c.f. section 7.2.1) which the PDTB parser successfully resolves using various structural features. The poor result of the Hyland feature is likely due to the dominance of the Argument stage across both polar and neutral sentences.

Aside from the PDTB feature set, the most dramatic accuracy gain across all three classifiers was achieved using a bag of stemmed, stance-generalized words. When we examined the model generated by the NBMB, we found that many of the terms that served as good indicators of the *polar* class included stancetaking verbs discussed in section 3.6 (*determine, develop, think*), along with DMs (*nevertheless, though*). The presence of stancetaking lexis and DMs in a sentence can therefore serve as a good indicator of that sentence's *polar* status. We also find several prompt topic-specific words in this group. Since writers will often make mention of the prompt topic when developing an SA backing a stance *for* or *against* a given essay prompt, prompt topic mentions such as *feminist, punish, rehabilitate*, and *technology* were often associated with *polar* sentences.

Table 9.5, which gives the 15 highest-weighted terms learned by the SVM neutral-polar classifier shows a similar pattern. Stancetaking terms (generalized to `STANCE_WORD`) and prompt topic words (*money, crime, imagine, television*, etc.) ranked highest among the weights learned for the *polar* class, though DMs were not as highly weighted as stancetaking terms and topic terms.

Polar ITEM WEIGHTS

LEXICAL ITEM	WEIGHT
dream	2.6335
comfort	1.7866
STANCE_WORD	1.6752
money	1.6452
crime	1.6367
imagin	1.5705
kill	1.4354
confid	1.4272
televis	1.3252
socialist	1.3232
forget	1.3205
student	1.2875
comput	1.2803
method	1.242
evil	1.2418

Neutral ITEM WEIGHTS

LEXICAL ITEM	WEIGHT
wide	-1.521
produc	-1.4248
done	-1.4099
longer	-1.3711
both	-1.3645
got	-1.2988
latest	-1.294
such	-1.2862
over	-1.2777
separ	-1.2336
babysitt	-1.1784
creation	-1.1775
shown	-1.1358
materi	-1.1327
vital	-1.1286

Table 9.5: Top 15 polar and neutral feature weights learned by the SVM neutral-polar classifier.

9.3 Polarity classification: Experiment setup and results

We used 1572 polar sentences in our polarity classification experiments, making sure that none of these sentences were previously used in the *neutral-polar* classification experiments. 883 (=56%) of these sentences were tagged *for* by annotators, while 689 were tagged *against* (=44%). All features used in the polarity classification experiments are given in Table 9.6. As shown in Table 9.7, we compared the performance of the three standard text classification algorithms also used in our *neutral-polar* classification experiments: NB, NBMB, and an SVM classifier with a linear kernel. As with our *neutral-polar* classification experiments, the significance of accuracy increases resulting from cumulatively adding features to the majority-tag model was measured using McNemar's χ^2 test at level $p< .05$. The 5-word PMI expansions and the 15-word expansions were each evaluated separately. In other words, a model consisting of opinion-generalized words, essay stance, and 5-word PMI expansion sets was evaluated and the resulting increase in accuracy was measured relative to the combined opinion-generalized words and essay stance feature sets.

FEATURE	DESCRIPTION
Opinion generalized unigrams	All sentences are transformed into unordered vectors of words, with each content word matched to the MPQA subjectivity lexicon generalized to `POS, NEG`, or `NEUT`. Each word is represented as a separate feature that is either frequency-valued (in the case of Naive Bayes and the SVM classifier) or binary-valued (in the case of Multivariate Bernouli Naive Bayes).
Essay stance	Nominal valued-feature indicating the gold stance polarity tag (*for* or *against*) of the essay from which the sentence was extracted.
5-word PMI expansion set	For each noun and verb in the sentence, the 5 highest-scoring words (determined using the PMI metric) extracted from Wikicorpus. Each word in the expansion set is represented as a separate feature that, depending on the classifier used, is either frequency-valued or binary-valued
15-word PMI expansion set	For each noun and verb in the sentence, the 15 highest-scoring words (determined using PMI) extracted from Wikicorpus. Each word in the expansion set is represented as a separate feature that is either frequency-valued or binary-valued

Table 9.6: Features used in the Polarity classification experiments.

This process was then repeated using the 15-word version of the PMI-scored expansion sets. For each classifier, only the best-performing PMI set is shown in Table 9.7.

9.3.1 Discussion

Table 9.7 shows the results of the polarity classification task. For each of the classifiers used in the polarity classification experiments, the highest-accuracy combination of features achieved an accuracy greater than a bag-of-words baseline. The highest accuracy overall was 79%, achieved using a NBMB classifier trained on a combination of opinion-generalized unigrams, essay stance tags, and 5-word PMI expansion sets. This score was significant relative to a bag-of-words baseline at level $p <.02$. The second highest overall accuracy was 77%, achieved using an SVM trained on opinion-generalized unigrams, essay stance tags, and 15-word PMI expansion sets. This boost

		for			*against*		
	Acc.	Prec.	Rec.	F	Prec.	Rec.	F
Naive Bayes							
Baseline: Bag of words	66.2	72.3	63.9	67.8	60.4	69.3	64.5
None (majority tag)	57.8	58.0	90.3	70.7	56.8	16.4	25.5
Opinion-gen. words	61.9*	67.4	62.4	64.8	56.0	61.4	58.6
Opinion-gen. words + essay stance	63.2	68.7	63.6	66.1	57.4	62.8	60.0
Opinion-gen. words + essay stance + 15 PMI exp.	**70.4***	72.5	76.3	74.4	67.5	63.0	65.2
Multivariate Bernouli Naive Bayes							
Baseline: Bag of words	76.8	78.3	80.8	79.5	74.8	71.9	73.3
None (majority tag)	57.8	58.0	90.3	70.7	56.8	16.4	25.5
Opinion-gen. words	68.3*	69.7	77.3	73.3	66.2	56.9	61.2
Opinion-gen. words + essay stance	73.9*	75.6	79.2	77.4	71.6	67.3	69.4
Opinion-gen. words + essay stance + 5 PMI exp.	**79.0**	79.3	84.8	81.9	78.6	71.6	74.9
SVM (linear kernel)							
Baseline: Bag of words	71.3	74.2	74.2	74.2	67.6	67.6	67.6
None (majority tag)	57.8	58.0	90.3	70.7	56.8	16.4	25.5
Opinion-gen. words	64.1*	67.5	69.9	68.7	59.6	56.9	58.2
Opinion-gen. words + essay stance	70.1*	73.8	72.5	73.1	65.5	67.1	663.
Opinion-gen. words + essay stance + 15 PMI exp.	**77.0***	78.7	81.0	79.8	74.7	72.0	73.3

Table 9.7: Polarity classification experiment results. For each classifier, the highest accuracy model is boldfaced. The highest accuracy model across all three classifiers is boldfaced and underlined. * indicates a statistically significant increase in accuracy at level $p < .05$ (using McNemar's test) relative to the feature set immediately above.

in accuracy was significant at level $p < .01$ relative to baseline. Finally, the NB classifier, trained on the same combination of features as the SVM classifier, achieved an accuracy 70.4%, which was significant at level $p < .001$ relative to baseline.

Examining the increases in accuracy resulting from cumulatively adding features to the majority tag model, we observe significant increases in accuracy for all feature sets for all but one of the classifiers. Adding the essay stance feature to the opinion-generalized words feature failed to significantly increase the accuracy of the NB classifier. In contrast to the *neutral-polar* classification models discussed in the last section, adding generalized lexical items, while helpful, was not as helpful as the combined addition of essay stance and PMI expansion features. The 15 highest-weighted *for* and *against* features learned by the SVM classifier are given in Table 9.8.

Recall our discussion in section 8.2.2 regarding the relationship between the opinion polarity of the language used in each prompt text and the polarity of an SA backing a stance towards that *for* FEATURE WEIGHTS

Feature	weight
reprisal	1.3864
diffraction	1.26
androids	1.2364
sucker	1.2159
escalates	1.1062
mistake	1.1053
fattest	1.0737
scrambled	1.0682
all	1.0597
who	1.0495
businessmen	1.0143
`essay_stance`	1.012
critical	1.0099
unbound	1.0039
degree	0.9768

against FEATURE WEIGHTS

Feature	weight
unblemished	-1.4826
coevolution	-1.2923
maximizing	-1.162
extroverted	-1.1281
wisely	-1.1184
conscience	-1.1126
knowledge	-1.0961
northwestward	-1.075
interchangeably	-1.061
inspirational	-1.06
jails	-1.0323
rainbows	-1.0246
satisfiability	-1.0015
bystanders	-1.0
prize	-0.9472

Table 9.8: Top 15 *for* and *against* feature weights learned by the SVM polarity classifier.

prompt text. In 8.2.2 , we hypothesized that opinion-bearing features could help to discriminate the stance polarity of an SA. This is due to the tendency of writers to align or disalign themselves with a prompt text by using opinion-bearing language that is either matched to the polarity of the prompt text (in the case of an aligned stance) or is of an opposite opinion polarity to the language in the prompt (in the case of disalginment). We find some support for this hypothesis in Table 9.8. Since much of the language in the seven ICLE prompts is negative, we expect to see negative opinion-bearing language (*reprisal, sucker, mistakes*) associated with high-weighted *for*-stanced features and positive language (*unblemished, extroverted, wisely*) associated with high-weighted *against*-stanced features.

The essay-stance feature proved a better indicator of *for* stance than *against* stance: SAs extracted from *for*-stanced essays were themselves usually *for*-stanced, though SAs extracted from *against*-stanced essays were less reliably *against*-stanced. We can observe this pattern in Table 9.8: `essay_stance` is ranked high in the list of feature weights associated with a *for* stance, but is not included in the highest features associated with an *against* stance.

PMI-expansion words, which captured common topics mentioned by the SAs, were often good discriminators of stance. Examining the NBMB model, we found that PMI expansions such as *coevolution, conscience, hallucinations* were highly-discriminative. Such words were also ranked high in the lists of *for* and *against* feature weights learned by the SVM classifier shown in Table 9.8. Interestingly, many of these terms are opinion-bearing and their inclusion in the lists of high-weighed SVM feature weights follows the pattern observed above for positive opinion-bearing terms. Terms such as *unblemished, extroverted, wisely* are ranked high in the list of *against* features, while negative opinion-bearing terms, such as *reprisal, sucker*, and *mistake*, were ranked high as *for* features. Again, this confirms our hypothesis that opinion-bearing language can serve as a good indicator of the stance polarity of an SA.

9.4 Summary

In this chapter, we described the construction of two SA classification models. The first, neutral-polar classification model identifies a given sentence as or as not an SA. The second, polarity classification model classifies an SA as holding a *for* or *against* stance. Our supervised approach to these tasks involved the construction of a corpus of 8176 sentences (extracted from 239 ICLE essays) annotated as *for*, *against*, or *neutral*. Crowdsourced annotation of these sentences resulted in observed agreement and κ scores of .85 and .70, respectively, between a set of crowdsourced annotations and a gold-standard set completed by the author.

In section 9.2, we described the results of a series of neutral-polar classification experiments using three standard text-classification algorithms—NB, NBMB, and SVMs. The lexical, discourse structure, and rhetorical structure features developed in chapter 8 were cumulatively added to a majority-tag model and the significance of any increase in accuracy was evaluated. We found that PDTB-based features and stance-generalized features make significant contributions to accuracy increases across all three classifiers, but features based on Hyland's Thesis-Argument-Conclusion scheme and an SA-lexicon lookup feature fails to make significant contributions. The highest overall accuracy was 74.3%. This accuracy was achieved using a NBMB model trained on all four feature sets and was significant at level $p < .01$ relative to an accuracy of 69.8% achieved using a NBMB model trained on a bag-of-words feature set.

A series of polarity classification experiments were reported in section 9.3. In these experiments, the three polarity classification features described in chapter 8 were used to train a NB, NBMB, and SVM classifier. As with the *neutral-polar* classification experiments, we cumulatively added each feature to the majority-tag model and evaluated the significance of any increases in accuracy. In contrast to the performance of the *neutral-polar* classification features, all three features used in the polarity classification experiments were found to make significant contributions to the model learned by each classifier. The highest overall accuracy was the 79.0% score achieved by a NBMB model trained on a combination of opinion-generalized words, essay stance tags, and 5-word PMI expansions. This score was significant at level $p < .02$ relative to a bag-of-words baseline accuracy of 76.8%.

Chapter 10

Application: Supporting Argument Summarization

This chapter introduces a new NLP task that incorporates many of the classification models presented in this study, *supporting argument summarization*. This task draws on two lines of research in Sentiment Analysis, *opinion reason mining* and *single-sentence opinion summarization*. As we describe in detail in section 8.1, opinion reason mining involves the automated identification and classification of opinion reasons in review text (Kim and Hovy, 2006; Branavan et al., 2009; Zirn et al., 2011; Sauper et al., 2011; Qin et al., 2008). Single-sentence opinion summarization, introduced in Glaser and Schütze (2012), involves the extraction of single-sentence summaries from product reviews that indicate the review author's overall opinion regarding the product and provides a good reason supporting that opinion. In our case, the goal is to extract a single sentence from a given essay that gives a reader a sense of the overall stance of the essay from which the sentence was extracted and also provides a good SA for that stance.

10.1 Background

As described in Mani (2001), there have historically been two approaches to automated text summarization, an *abstractive* approach and an *extractive* approach. The abstractive approach uses topic-modeling and language generation techniques to create well-formed text summarizes of one or more documents that, ideally, should be indistinguishable from human-authored summaries. Abstractive approaches tend to be template-driven with (optionally modified) salient phrases first extracted from raw text and then inserted into hand-constructed template slots. The SUMMONS news summarization system of McKeown and Radev (1995), for example, first gathers salient information from single or multiple documents via a *content planning* component containing various heuristic rules designed to recognize and resolve multi-sentence relationships such as *change of perspective* and *contradiction*. The filled-template version of this content is then handed to a *linguistic generator* that selects a presentation format for the extracted information.

More commonly, however, researchers have taken an extractive approach to text summarization, an approach that dates to the earliest work in the field (Luhn, 1958; Edmunson, 1969). Extractive text summarization involves identifying salient sections of the target text and returning them to the reader in their original form. In the case of opinion summarization, these sections tend to be one or more key sentences that capture a document's overall opinion toward some entity (Stoyanov and Cardie, 2006a,b) or toward particular aspects of that entity (Hu and Liu, 2004; Titov and McDonald, 2008).

While a typical aspect-based opinion summarization system might accurately return a group of opinions regarding one or more aspects of a target entity such as a gadget ("I love the user interface") or movie ("I hated the ending"), such systems provide little information as to *why* writers feel the way they do about these aspects. Why, for example, does a writer of a review of a phone dislike the phone's interface or appearance? Kim and Hovy (2006) address this point in the context of an aspect-based summarization task dealing with product reviews. They introduced a new extractive summarization task, *opinion reason identification*. The goal of this task is to extract a single reason per review that both captures the writer's overall opinion and presents a a good reason in support of that opinion. Many of the details of Kim and Hovy's system were described in section 8.1, however, we will briefly review their methodology. Kim and Hovy first collected ~300,000 sentences from two online product review sites and semi-automatically labeled these sentences as *pro*, *con*, or *neither* (for Kim and Hovy, *pro* and *con* refer to reasons supporting a positive or negative opinion, respectively). A two-stage opinion reason classification model was constructed. The first, *neutral-polar* classification model (is the sentence a reason?) achieved its highest accuracy of 76.2% on a test set of mp3 player reviews, and the second, *polarity* classification model (is the reason identified in the first stage *pro* or *con*?) achieved its highest accuracy of 57.1% also on a set of mp3 player reviews.

The opinion reason mining work most similar to the SA summarization system presented in this chapter is the "single-sentence summarization" system described in Glaser and Schütze (2012). In section 8.1, we discussed Glaser and Schütze's basic approach to the generation of *supporting sentences* that summarize both the positive or negative opinion associated with a review and the writer's reason for that opinion. Inspired in part by Glaser and Schütze's research, we introduce a single-sentence summarization task for argumentative text, *supporting argument summarization*, which involves the extraction of a single SA per document that conveys the central reason supporting the author's overall argument stance. To evaluate SAs generated by our system we adopt the novel evaluation method introduced in Glaser and Schütze's study, as described in section 10.4.

10.2 Data

Our first task was the creation of a test set of SAs. The SA annotation work described in section 9.1 involved the annotation of a 239-essay subset of our original set of 1320 essays. After subtracting this subset, this left us with 1081 essays for our SA summarization experiments. We

ESSAY PROMPT	#ESSAYS	%FOR	%AGAINST
Armies	46	.80	.20
Degrees	100	.52	.48
Feminism	48	.31	.69
Marx	61	.85	.15
Money	38	.45	.55
Prisons	50	.72	.28
Science	159	.29	.71
Total	502	.51	.49

Table 10.1: Distribution of essay prompts and stance polarities for 502 essays used in the supporting argument summarization task. Stance polarities are taken from the essay-level stance annotation task described in section 4.3.

examined this 1081-essay subset and extracted those essays that struck us as the most proficiently written, in terms of structure, diction, grammar, and coherence of argumentation. These essays, we hypothesized, would be more likely to contain coherent supporting arguments of the sort that could be reliably evaluated by our crowdsourced raters. This process left us with 502 essays. The distribution of these essays relative to each prompt is given in table 10.1.

10.3 Method

The first part of our supporting argument summarizer is built in a two-stage or "cascade" fashion. The first-stage, *neutral-polar* classifier is used to identify any sentences in an essay that are SA's, and a second-stage, polarity classifier is used to classify SAs generated by the first-stage classifier as containing a *for* or *against* polarity. We find several precedents for two-stage sentiment classification systems in the literature, the most relevant of which are the systems of Pang and Lee (2004), Kim and Hovy (2006) (described in detail in section 8.1), and Wilson et al. (2009). Pang and Lee (2004) is an early example of a two-stage system for the classification of sentiment polarity in movie reviews. Their system first reduces a given review to an "excerpt" containing subjective sentences and then classifies these sentences as positive or negative. More recently, Wilson et al. (2009) described a two-stage sentence-level sentiment classification system designed to resolve contextual polarity ambiguity at the phrase and sentence level by first identifying all polar text segments and then classifying these segments as positive or negative.

Our SA summarization system is built in four steps, with the first two steps corresponding to stages 1 and 2 of our SA stance classifier. The third step involves the extraction of the highest-ranked *for* and *against* SAs. The fourth and final step involves the use of oracle information provided by the annotated stance of the essay from which the SAs were extracted to pick the highest-ranked SA that displays a stance polarity matched to the essay's stance polarity. These steps are summarized below.

- **Step 1:** The first-stage classifier is the highest-accuracy *neutral-polar* classifier described in section 9.2 (a NBMB model trained on all four *neutral-polar* classification feature sets) which is used to identify any *polar* (=SA) sentences in a given essay.

- **Step 2:** The second-stage classifier is the highest-accuracy polarity classifier described in section 9.3 (a NBMB model trained on all three feature sets, including a five-word PMI set feature) which is used to classify the *polar* sentences returned by Step 1 as arguments in support of a *for* or *against* stance.

- **Step 3:** When making its predictions, the second-stage classifier assigns each sentence a probability of class *for* or class *against*, with the higher of these probabilities determining the class assigned to the sentence. In Step 3, we use these probabilities to create two rank-ordered lists of sentences assigned a *for* polarity and sentences assigned an *against* polarity.

- **Step 4:** The final step involves choosing which of the two highest-ranked sentences output by Step 3—the highest-ranked *for* SA or the highest-ranked *against* SA—should be chosen as the representative SA of that essay. For this step, we use oracle information provided by the overall argument stance assigned to the essay by our essay-level stance annotators (cf.

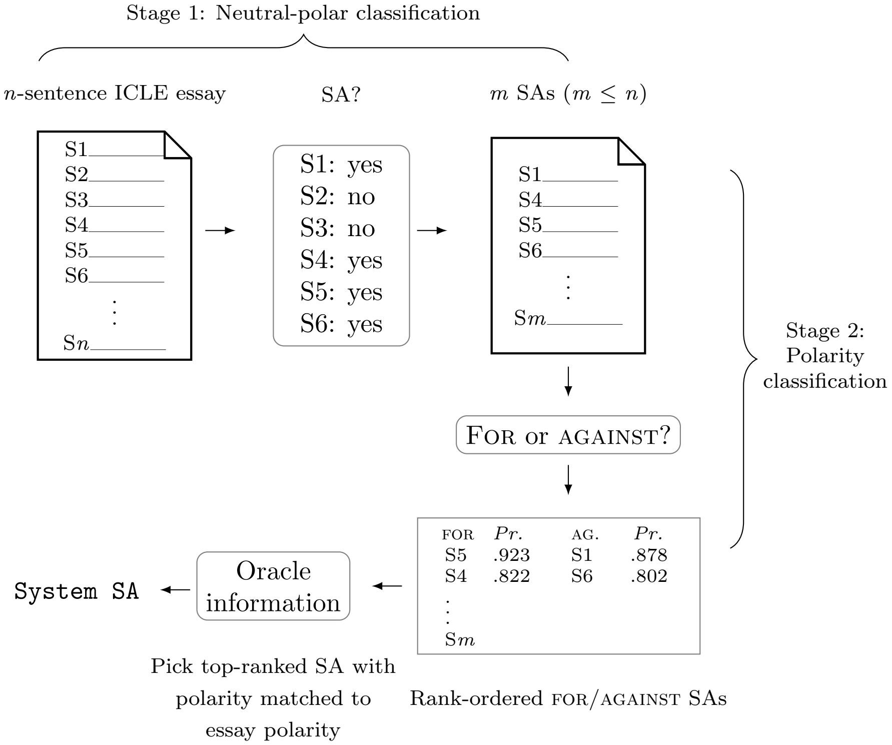


Figure 10.1: Flow of summarization system.

chapter 4) to extract the highest-ranked SA that also matches the stance polarity assigned to that essay.

These steps are summarized in the diagram given in Figure 10.1.

Examples of SAs generated using the system outlined in Figure 10.1 are given in Table 10.2. For each of our seven prompts, this table shows the SAs extracted from two essays responding to that prompt (one *for* and one *against*).

We can make a number of observations regarding the SAs given in Table 10.2. First, most of the SAs returned by our system are not fragments of larger arguments and do not require the original essay context to resolve any ambiguities regarding their interpretation. After reading each SA, the reader has a notion of the likely stance of the text from which the SA was extracted and does not require any other sentences from the text to understand the argument being made. This

PROMPT	SYSTEM SA	*For/Against*
Armies	I think that the decision to use only professional soldiers is a good one because they know what they do and they do it out of their own free will.	*for*
	Finally, everybody thinks that military service is the best and only way for a teenager to become a man, that is to say, to mature.	*against*
Degrees	If the students are prepared theoretical during university education they might have difficulties in putting these theories into the practice in the real world.	*for*
	Even it is theoretical, we need it because it is the basic for us to develop our knowledge, and also our career.	*against*
Feminism	Criticisms are one thing, but a central flaw in the evolution of women's liberation is that women try to become superwomen, to prove that it is possible to match men in the workplace as well as intellectually while also fulfilling the role of mother	*for*
	Atlhough there have been few changes which we can remember these last years, we really should take our hats off to feminists: they have been challenging a mentality established since the beginnings of the world and for a great part improving it	*against*
Marx	For some people television seems to offer too much, so much that almost everything else in their life becomes less and less interesting and significant - they stop living their own life and start living in this unreal TV-world.	*for*
	It is a great, useful invention; it allows us to know what happens in the world everyday; through it you can see place that you will never visit, cultures that you don't know because they are far and different from yours.	*against*
Money	Sure, money is the root of all the evil because it causes killing of people, it is a great mean of power, it divides people into different social classes.	*for*
	And isn't it a paradox that money which is considered to be the root of all evil serves as the root of good.	*against*
Prisons	Criminals who sit for many years in a cell with no one to talk to have probably not become better persons.	*for*
	But on the other hand, if you have committed a crime, you have broken a rule, you have in one way no right to live among civilised people, and you have no one to blame but yourself.	*against*
Science	How many people would prefer a novel in a book where they have to think, they need a lot of time to create their own opinion and imagination, to a novel on tv where everything is served in a few minutes?	*for*
	Science technology adds to our imagination, and other way around, there would not be scientific discoveries if there were not ”crazy scientists” with vivid imagination and big dreams.	*against*

Table 10.2: *For* and *against* system SAs extracted from 14 ICLE essays.

is an encouraging result, since the goal of our SA summarization system is to provide the reader with an at-a-glance assessment of both the stance of the original essay and the most representative argument used to support that stance in a single sentence.

Second, our system captures many of the more prevalent arguments supporting or opposing propositions dealing with such controversial issues such as the legacy of feminism and the social impact of science and technology. As noted in work dealing with automated argument categorization in healthcare reform debate data (Conrad et al., 2012), reasons given in support of *for* or *against* arguments regarding highly divisive socio-political issues tend to fall into predictable classes. Conrad et al. found that eight *for* reasons and eight *against* reasons could reliably capture all of the various arguments *for* or *against* healthcare reform. For example, common *for* SAs related to healthcare reform include *controls healthcare costs* and *helps economy*, while common *against* SAs include *expands government* and *hurts economy*. Similarly, student writers arguing for the claim that feminism has done more harm than good often draw attention to a purported conflict between parenting and career responsibilities, as we see in the *for*-stanced SA given in response to the *Feminism* prompt in Table 10.2. We find another common strand of argumentation in the *against*-stanced SA responding to the *Science* prompt in Table 10.2. When arguing against the claim that science and technology have displaced imagination, the writer argues that progress in science and technology is contingent upon the existence of imagination. Manual examination of the *Science* section of the ICLE corpus revealed this to be a common *against* argument.

10.4 Evaluation

As Table 10.2 shows, our system generates interesting SAs that can serve as single-sentence summaries of the overall stance of the essay from which the SA was extracted and the reason given in support of that stance. Evaluating this output is a somewhat more complicated task than evaluating document- or sentence-level classifier output. Generally, evaluating a system-generated extractive summary presumes the existence of human-generated gold-standard summaries. System-generated summaries are then scored relative to the gold-standard summaries using lexical overlap metrics, such as the cosine similarity, ngram overlap, or longest-common-subsequence metrics proposed in Saggion et al. (2002), or ngram co-occurrence metrics such as the ROUGE metric of Lin (2004). While sentiment annotation tasks of the kind described in sections 4.3 and 9.1.2 require minimal training and can be reliably entrusted to crowdsourced annotators, generating gold-standard extractive summaries requires significant training and, occasionally, expert knowledge. These requirements can be prohibitive for many researchers. As noted in Lin (2004), for example, the annotation effort required to evaluate a typical large-scale summarization system would require over 3,000 human effort hours since annotators must first read the document to be summarized in its entirety and thereafter, depending on the summarization task at hand, extract one or more sentences that best capture some feature of that document.

In the case of single-sentence summarization, a crowdsourced approach to system evaluation, as introduced in in Glaser and Schütze, can serve as a cost-efficient alternative to an expert annotation effort. This approach involves pairing all sentences generated by the summarization system with baseline sentences and then asking annotators to pick which sentence offers a more "convincing reason" in support of a given opinion. The accuracy of the system can be evaluated by noting the percentage of sentences that were adjudged more convincing reasons than their baseline counterparts. Glaser and Schütze used this method to evaluate the single-sentence opinion reason summaries of 1380 product reviews. Each of the 1380 opinion reasons generated by their system (one for each review) was paired with a baseline sentence. All baseline sentences were generated by identifying the opinion polarity of all sentences in the review and extracting the single sentence that was identified with the highest confidence by a sentence-level opinion classifier. Crowdsourced raters consider the system sentence generated by Glaser and Schütze's system a more convincing reason in 64.6% of cases.

In replicating Glaser and Schütze's evaluation method, the first step was the construction of a set of suitable baseline sentences. This necessitated the creation of a sentence-level stance classification counterpart to the sentence-level opinion classifier used by Glaser and Schütze to create their baseline sentences. For this step, we adopted the sentence-level stance classification component of the document-level classification system presented in Somasundaran (2010) and Somasundaran and Wiebe (2010). As described in section 5.2.5, one important component of Somasundaran and Somasundaran and Wiebe's document-level stance classification system involved identifying the dominant stance polarity of sentences by summing the probabilities associated with all *for* and *against* stancetaking words found in each sentence and then assigning the higher scoring polarity to each sentence. We replicated this procedure for all 502 essays used in the SA summarization task using the stancetaking words contained in the modified stance lexicon described in section 5.2.5. After determining the dominant stance and score for each sentence we extracted the two highest-scoring *for* and *against* sentences from each essay. We then chose the highest-scoring stancetaking sentence whose polarity matched the document-level stance polarity of the essay from which the sentence was extracted (the document-level stance polarity was determined using the document-level annotations described in section 4.3). This sentence served as that essay's baseline sentence.

Typical system sentence/baseline sentence pairs are given below as (105–107). Each pair contains a system sentence given in Table 10.2 and is paired with the baseline sentence identified by the sentence-level stance classifier. One can see that, in contrast to the SAs generated by our system given in Table 10.2, the baseline sentences in (105–107) are generally fragmentary, do not convey the overall stance of the essays from which they were extracted, and do not offer coherent reasons in support of a *for* or *against* stance.

(105) FEMINISM PROMPT/AGAINST

a. SYSTEM SA: Criticisms are one thing, but a central flaw in the evolution of women's liberation is that women try to become superwomen, to prove that it is possible to match men in the workplace as well as intellectually while also fulfilling the role of mother.

b. BASELINE SENTENCE: If women are not freed from these functions they will never achieve freedom and therefore artificial reproduction is urged as a means of freeing women.

(106) PRISONS PROMPT/FOR

a. SYSTEM SA: Criminals who sit for many years in a cell with no one to talk to have probably not become better persons.

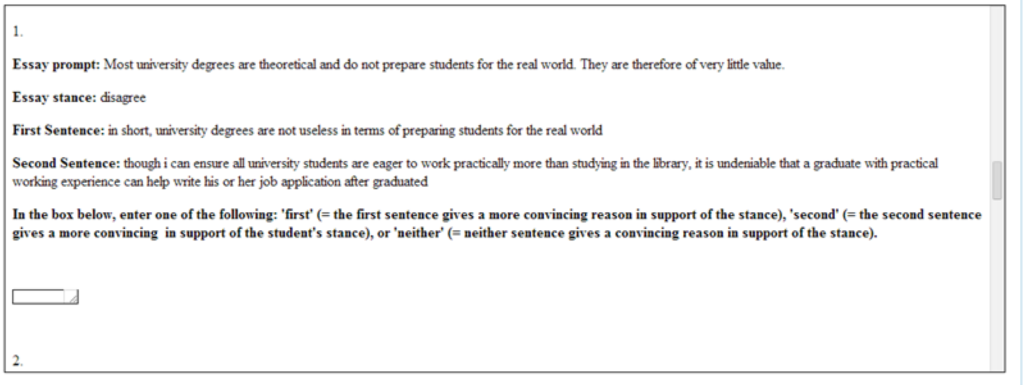
b. BASELINE SENTENCE: In the prisons, there are a clear distinction on how the criminals are treated, and this depends on whether the crime is serious or not.

(107) DEGREES PROMPT/FOR

a. SYSTEM SA: If the students are prepared theoretical during university education they might have difficulties in putting these theories into the practice in the real world.

b. BASELINE SENTENCE: The more important thing is that they are going to save their lives to have a better future.

The system sentence and baseline sentence were identical for 34 of 502 essays. These pairs were not used in the evaluation tasks reported here. The remaining 468 system SA/baseline sentence pairs—one pair for each of the 468 essays—were posted to AMT and three unique annotators were assigned to each pair. Annotators were given the essay prompt associated with the essay from which each sentence pair was extracted, the stance of the essay from which the pair was extracted (provided by the essay-level stance annotation work described in section 4.3), and the (randomly ordered) system SA/baseline sentence pair associated with that essay. Annotators were asked to identify which of the two sentences gave a more convincing reason in support of the provided stance by writing the word “first” or “second” in a text box. If neither sentence gave a convincing reason, annotators had the option of writing “neither” in the text box (the full annotation protocol used for this task is provided in Appendix C) . As with the other annotation tasks described in this study, the selection of annotators was restricted to those with U.S.-based IP addresses and high annotation acceptance rates. Figure 10.2 shows the AMT interface used by evaluators. To determine the final annotation of each sentence pair, we adopted the scoring system of Glaser and Schütze. This system counts the number of times a given sentence is rated better than its competitor sentence. Since there are three raters, each sentence receives a score of 0,1,2, or 3, with the



1.

Essay prompt: Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.

Essay stance: disagree

First Sentence: in short, university degrees are not useless in terms of preparing students for the real world

Second Sentence: though i can ensure all university students are eager to work practically more than studying in the library, it is undeniable that a graduate with practical working experience can help write his or her job application after graduated

In the box below, enter one of the following: 'first' (= the first sentence gives a more convincing reason in support of the stance), 'second' (= the second sentence gives a more convincing in support of the student's stance), or 'neither' (= neither sentence gives a convincing reason in support of the stance).

2.

Figure 10.2: Screenshot of the AMT interface for the SA summarization evaluation task.

higher-scoring sentence considered a better SA than its competitor.

To evaluate agreement we used Fleiss' κ, which measures nominal-scale agreement—adjusted for chance—between *n*-numbered annotators. Fleiss's κ is generally defined as

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}},$$

where the numerator $\overline{P} - \overline{P_e}$ is the proportion of agreement actually achieved above chance and the denominator $1 - \overline{P_e}$ is the proportion of agreement that could potentially be achieved above chance. Fleiss' κ for this task was .52 (interpreted as "moderate" agreement by Landis and Koch, 1977) while observed agreement was .70. Of the 468 sentence pairs evaluated, annotators preferred the system sentence in 269 cases and the baseline sentence in 121 cases. 78 cases either resulted in a tie or were unanimously tagged *neither* by all three raters. Following Glaser and Schütze, we designate such pairs as *no decision*. To resolve as many of these pairs as possible, all 78 *no decision* pairs were re-posted to AMT and new annotators were assigned to evaluate each pair. This succeeded in resolving 34 of the 78 *no decision* pairs. The results of both AMT evaluation passes are given in rows 1 and 2 of Table 10.3. Row 3 of Table 10.3 consolidates the results of both passes.

As rows 1 and 3 of Table 10.3 show, raters clearly preferred the system-generated SA to the

	Experiment	#essays	Picked system		Picked baseline		No decision	
			%	#	%	#	%	#
1	AMT, first pass	468	57.4	269	25.9	121	16.7	78
2	AMT, second pass	78	23.1	18	20.5	16	56.4	44
3	AMT, both	468	**61.3**	287	29.3	137	9.4	44

Table 10.3: Summarization system evaluation results.

baseline SA in the majority of cases. In the first pass, raters preferred the system sentence in 57.4% of cases—a somewhat low percentage, but still well above baseline. After resolution of 34 of the ties produced in the first pass, the percentage of cases in which raters preferred the system SA increased to 61.3%.

To determine the significance of this result, we performed a binomial test on those cases in which the system sentence was preferred to the baseline sentence. After subtracting the 44 *no decision* pairs, the total number of cases is 424. As shown in the last row of Table 10.3, in 287 of these cases raters preferred the system sentence and in 137 cases they preferred the baseline sentence. Using the binomial test, raters' preference for the system-generated SA's was significant at level $p < .001$, two-tailed.

10.5 Error analysis

In this section, we examine cases in which raters preferred the baseline sentence to the system sentence. We will also take a look at examples of unresolved *no decision* pairs. Raters unanimously preferred the baseline sentence to the system SA under two scenarios. In the first scenario, a mismatch existed between the system SA polarity and the essay-level polarity, i.e., an SA purportedly given in support of the provided essay stance actually contained a better argument for an opposite-polarity essay stance. In the second scenario, the baseline sentence genuinely offered a more convincing reason than the system SA. In examples (108) and (109), we find that the system sentences provide generally convincing reasons in support of a *for* stance toward the *Science* and

Money prompts, respectively, yet the essays from which these SAs were extracted argued *against* the statements in these prompts. Raters therefore chose the baseline sentence as the better reason in these cases.

(108) SCIENCE PROMPT/AGAINST

a. SYSTEM SA: This means that those who work in this field have become victims of the technology— the technology controls the human being, the man has become a slave with no imagination.

b. BASELINE SENTENCE: It could be that they find no interest in the information technology or the technology as such, they have another dream or perhaps not a dream at all or they see that the society will not be in balance if everybody goes in the same direction.

(109) MONEY PROMPT/AGAINST

a. SYSTEM SA: As we can see, money can do a lot of harm and in this sense the popular saying 'money is the root of all evil' seems to be justified

b. BASELINE SENTENCE: For some it is the opportunity to travel all over the world, for others buying house or paying for their studies.

Other cases in which raters showed a preference for the baseline sentence involved sentences that in fact offered a more convincing reason than the system sentence. In (110), the system sentence given in (110a) appears to be a fragment of a larger argument while the baseline sentence given in (110b) provides a good reason in support of a *for* stance in the form of several pieces of evidence purporting to show that the legacy of feminism has been negative. The system sentence given in (111a) provides no apparent reason either *for* or *against* the *Science* prompt. By contrast, the baseline sentence (111b) describes a scenario in which science and technology "transform us into mechanically-thinking creatures," a scenario compatible with the *Science* prompt's claim that science and technology have displaced imagination and therefore a good reason in support of a *for* stance.

(110) FEMINISM PROMPT/FOR

a. SYSTEM SA: Nevertheless, the result of their ability to overcome obstacles was rather mediocre inasmuch as they were only given, for instance, after the first world war, the right to vote.

b. BASELINE SENTENCE: Feminists eventually distorted the image they wanted to project onto the public: they acquired a status bordering on superiority over men, which alienated most people from feminism, even if this movement managed to restore dignity and pride.

(111) SCIENCE PROMPT/FOR

a. SYSTEM SA: I live under a constant stress because I have to study for difficult exams all the time as well as attending lectures and seminars every day.

b. BASELINE SENTENCE: We may realize that they destroy something of our human nature, that they make us turn into mechanically- thinking creatures but we cannot prevent this process from advancing.

Nearly all of the unresolved *no decision* pairs involved sentences unanimously tagged *neither* by raters. In the *no decision* pairs given in (112) and (113), we find system and baseline sentences that are either fragments, as in (112a) and (112b), bare assertions, as in (113a), or are simply incoherent, as in (113b).

(112) SCIENCE PROMPT/FOR

a. SYSTEM SA: And I merely envy these people because they will live nearer to the nature, will gain satisfaction from everyday life, will lead a healthier life than us and will have relationships with each other based on understanding, patience, love.

b. BASELINE SENTENCE: Let's hope that this dream comes true and that the balance between the spiritual and material aspects of human nature be restored for the well-being of us all.

(113) DEGREES PROMPT/FOR

a. SYSTEM SA: Thus, the knowledge given in universities should be permanent in order to be used not only in education process but also in real life.

b. BASELINE SENTENCE: It is obvious that these universities demand extending their proficiency in practicality.

10.6 Summary

In this chapter, we presented one possible application for the complete set of classification models presented in this study: an SA summarization system that returns a single sentence that both captures the overall argument of the essay from which it was extracted and also provides a representative argument in support of that stance. Section 10.3 describes the process used to construct our summarization model. Using a 502-essay subset of the annotated set of ICLE essays described in section chapter 4 as a test set, our model was constructed in four steps. The first step utilized the *neutral-polar* classifier described in section 9.2 to identify any SAs in a target essay. In steps 2 and 3, these SAs were passed to the polarity classification model described in section 9.3 which returns a rank-ordered list of *for* and *against* SAs. Finally, in Step 4, we used oracle information provided by the essay-level stance annotation tag to pick the highest-ranked essay with a polarity matching that of the essay from which it was extracted.

In section 10.4, we address the challenge of evaluating the output of our SA summarization system. We first note that standard summarization evaluation methodologies usually involve substantial expense and annotation effort since they require the creation of gold-standard summaries. The creation of such summaries is too difficult a task to be delegated to non-expert, crowdsourced annotators. We describe an alternative summarization evaluation technique introduced in Glaser and Schütze (2012), which is particularly well-suited to single-sentence summarization evaluation and can be carried out using non-expert raters. This technique involves the creation of a set of baseline sentences—one for each of the system sentences we are interested in evaluating—which are then comparatively evaluated by raters. Our baseline sentences were created using an off-the-shelf sentence-level stance classifier described in Somasundaran (2010) and Somasundaran and Wiebe (2010). For each essay in a 502-essay subset of ICLE essays, the highest-ranked *for* and *against* sentences returned by this classifier were extracted. The sentence with a polarity that matched the document-level polarity of the essay from which that sentence was extracted was retained as that essay's baseline summary sentence. 468 of 502 SA/baseline sentence pairs were posted to AMT. After two annotation passes, our final results showed that raters chose the system sentence in 61.3% of cases, the baseline sentence in 29.3% of cases, and were undecided in the remaining 9.4% of cases. Raters' preference for the system sentence was significant at level $p < .001$ (using the binomial test).

Chapter 11

Conclusion and future work

This study has made four major contributions to the burgeoning field of stance classification. First, in place of the noisy online debate data currently used in document-level stance classification research, we compiled a corpus of quintessentially argumentative text, student argumentative essays. 1320 essays responding to seven very different prompt statements were annotated for document-level *for* and *against* stance. Second, we presented several high-accuracy document-level stance classification models trained and tested on this corpus. Using a set of features motivated by linguistic research involving stancetaking language, our best-performing classification model achieved an accuracy of 83% using the SVM learning algorithm. This accuracy was significantly higher than two baseline models. Third, we introduced the tasks of *supporting argument annotation* and *supporting argument classification*. After annotating 8176 essay sentences for supporting argument polarity, we constructed two sentence-level classification models. The first, *neutral-polar* classifier achieved an accuracy of 74.4%, and was designed to classify a given sentence as or as not a supporting argument. The second, *polarity* classifier achieved an accuracy of 79% for the task of classifying a supporting argument as providing a reason in support of a *for* or *against* stance toward a given prompt. Our final contribution introduced a novel summarization task, *supporting argument summarization*. The goal of this task is to extract a single sentence from a piece of argumentative text that indicates the overall argument stance of that text and gives a good reason for that stance. We designed and tested a proof-of-concept summarizer for a 502-essay subset of our student essays corpus. This summarizer utilized all three classification models and was evaluated relative to a baseline summarizer that extracted a single stancetaking sentence from an essay. Our crowdsourced evaluators preferred the system to the baseline sentence in 61.3% of cases, a result that was significant at level $p < .001$, two-tailed.

The following chapter outlines each of these contributions and describes several directions for future research.

11.1 Essay-level stance annotation and classification

Current stance classification research relies on data scraped from online debate forums, which are often noisy and unrepresentative of stancetaking text. One goal of this study was to test the effectiveness of classification features based on linguistic research involving argumentative language. To accomplish this, we required text that is representative of the kind of stancetaking language described in this research. We extracted 1320 essays from the ICLE corpus of Granger (2003), a collection of argumentative essays written by international students of English. These 1320 essays maintain the argumentative scenario described by researchers: a proposition is given and speakers/writers are asked to argue *for* or *against* that proposition. In chapter 4, we described an essay-level stance annotation effort, which involved the crowdsourced annotation of all 1320 essays for stance. This effort resulted in observed agreement and κ scores of .82 and .68, respectively.

Existing work involving document-level classification of stance has produced models achieving accuracies of of 63% for non-domain-specific online debate data (Somasundaran and Wiebe, 2010), and 69-74% for domain-specific debate data (Anand et al., 2011; Hasan and Ng, 2013a,b). We would argue that these somewhat low accuracy, non-domain-general scores are the result of a reliance on corpora that are unrepresentative of stancetaking text and feature sets that do not incorporate linguistically fine-grained information. Our goal in chapters 3-5 was to construct classification features entirely motivated by linguistic research involving stancetaking language. In chapter 3, we gave an overview of many of the more influential linguistic observations made regarding this register of persuasive language. Perhaps the most important of these observations concerned the semantic class of the targets of stancetaking language. Unlike opinion-bearing language, which targets entities such as movies and books and evaluates them positively or negatively, stancetaking language targets whole propositions and argues *for* or *against* their truth or likelihood. Lexico-syntactically, this difference is realized in the kind of lexical items associated with opinion-bearing versus stancetaking language. Opinion-bearing language tends to be adjectival, since its role is to evaluate nominal elements, while stancetaking language tends to be modal or evidential, since its role is to comment on the truth or likelihood of material contained in an accompanying clause.

In chapter 5, we described the construction of a set of classification features that incorporate these linguistic observations. To capture evidential word occurrence in stancetaking language, we used an adapted version of the lexicon of stancetaking words constructed in Somasundaran and Wiebe (2010). We also suggested that information regarding proposition targets of stance could be captured by locating any opinion-bearing words in the clause that accompanies a stancetaking word. We then introduced a novel feature type: the POS-generalized stance-attitude dependency subtree. This feature consists of a subtree extracted from dependency-parse representations of each essay sentence, with the head of the tree matched to our stance lexicon and the tail matched to a lexicon of opinion-bearing words (Wilson and Wiebe, 2005). A second feature set consisted of essay words that were semantically similar (as determined by the Wikipedia Link-based Measure of Witten and Milne, 2008) to words in the prompt to which the essay is responding. An SVM classifier trained on these feature sets achieved an accuracy of 82% which was greater than two high baselines.

11.2 Annotation and classification of supporting arguments

Knowing the argument stance of a document is rarely the only goal of either readers or systems of argumentative text analysis. We are also generally interested in the reasons, or *supporting arguments* the writer provides in support of their stance. In chapter 7, we looked at several linguistic and philosophical descriptions of supporting arguments and, in chapter 8, we incorporated the central strands of these descriptions into a set of supporting argument classification features. These features include occurrence of stancetaking and opinion-bearing words, discourse structure tags, and rhetorical/organizational structure tags. To test the effectiveness of these features, we collected crowdsourced annotations for 8176 sentences extracted from a sub-set of ICLE essays. Observed and κ interannotator agreement scores for this task were .85 and .70, respectively. Two sentence-level classification models were trained and tested on these sentences: a *neutral-polar* classifier, which identifies a sentence as or as not a supporting argument, and a *polarity* classifier which classifies the argument stance of a supporting argument as *for* or *against*. Accuracies for these two classifiers were 74.0% and 79.0%, respectively.

The supporting argument summarizer described in chapter 10 utilizes our supporting argument classification models to identify any supporting arguments in an essay and to classify these supporting arguments as containing a reason in support of a *for* or *against* stance. The overall stance of the essay from which the classified supporting arguments were extracted serves as oracle information guiding our selection of the highest-ranked supporting argument that has a stance polarity matching the essay's stance polarity (rank is determined by probability scores used by the polarity classifier to determine the polarity of the supporting argument). To evaluate our summarizer, we adopted the comparative evaluation approach of Glaser and Schütze (2012). This involved extracting two sentences from each of 502 ICLE essays—the supporting argument extracted by our summarization system and the highest-ranked stancetaking sentence (as determined by an off-the-shelf sentence-level stance classifier). These sentence pairs were posted to AMT and annotators were asked to determine which sentence provided the more “convincing reason” in support of a given essay’s stance. Our final results were encouraging: in 61.3% of cases, annotators preferred our system’s sentence to the baseline sentence.

11.3 Directions for future research

This study has laid the groundwork for several new tasks in stance classification research. In the area of stance annotation, we have shown that even non-expert annotators have sophisticated intuitions regarding both document-level and sentence-level stance. Although ICLE’s copyright restrictions prevent us from releasing a full-text version of our annotated subsection of the ICLE corpus to the public, we plan to release a table of ICLE file IDs with their associated annotations. An important next step in this research area will involve the collection and annotation of argumentative text comparable to the ICLE data and the release of this data to researchers. Additionally, more reliable annotations could conceivably be collected from expert rather than crowdsourced annotators.

Our document-level stance classification results suggest that linguistically motivated features provide measurable improvements in accuracy relative to less linguistically informed approaches. However, it is still not clear which linguistic aspects of these features contribute to the sizable accuracy increases observed in our document-level classification experiments. Do stancetaking words drive the performance of our dependency subtree feature? If so, which classes of stancetaking words? Modal verbs, epistemic judgment verbs, or modal adverbs? Answering these questions could provide us with interesting insights into the still poorly understood phenomenon of stancetaking language. Further, the performance of our classifier could potentially improve with selective pruning of less informative feature classes.

In the area of supporting argument classification, there are several interesting research avenues that have yet to be explored. First, our experiments with rhetorical and organizational features produced a somewhat surprising result: an organizational feature (adapted from Hyland, 1990), indicating the sentence's position in one of three high-level organizational stages, failed to contribute to classifier accuracy. This result is counterintuitive since, in a typical argumentative essay, supporting arguments seem to regularly occur in the Argument, rather than Thesis or Conclusion stages. Thus, one would expect that identifying the organizational stage in which a sentence occurs would serve as a good indicator of that sentence's status as a supporting argument. The poor showing of this feature in our experiments does not mean that organizational information is not informative of a sentence's argument status. There could be any number of reasons for our negative result. The feature representation itself is a potential culprit. An interesting line of research potentially involves comparing the performance of our nominal-valued feature representation to a representation that encodes sentence position on a continuous scale. Or perhaps it would be more useful to encode lower-level organizational elements of Hyland's scheme—higher-level organizational elements may simply be too coarse-grained for this task.

While our proof-of-concept supporting argument summarizer performed well, there is room for improvement. As our error analysis of evaluation results showed, many of the sentences returned by our system as the best supporting argument for that particular essay contain stancetaking language matching the polarity of the essay stance, but do not contain a clear reason in support of that stance. This was not unexpected since a key feature of the first-stage classifier used in our system involves generalizing any stancetaking language found in a given sentence to the term `STANCE_WORD`. A potentially interesting research avenue would be to adjust the weight of sentences identified as supporting arguments by the first-stage classifier so that our second-stage classifier shows less preference for sentences containing high levels of stancetaking language. This may have the effect of pruning sentences that contain highly stanced language but lack most of the features of supporting arguments.

Appendix A

Essay-level stance annotation protocol

A.1 Background

This annotation task involves annotating essays according to their overall argument stance toward a given statement. You will be given an essay prompt followed by a student essay arguing *for* or *against* the statement in the prompt. Your task is to determine if the essay is

- arguing *for* the statement,
- arguing *against* the statement, or
- is *neither* arguing *for* nor *against* the statement.

A.2 Annotation format

You will be given an essay prompt followed by an essay written in response to the statement in the prompt. When you have determined the essay stance, check one, and only one, of the following:

- This essay is arguing *for* the statement in the prompt.
- This essay is arguing *against* the statement in the prompt.
- This essay is *neither* arguing *for* nor *against* the statement in the prompt.

A.3 Tips

- Trust your immediate impression of the essay and try not to overthink your annotations.
- You will encounter many spelling mistakes and grammatical errors in these essays. Try to ignore these errors and do your best to infer what the author is trying to say.

- Often there will not be an explicit *for/against* statement in the essay. Instead, the essay just gives a general impression of a *for* or *against* stance. This "general impression" is the stance of the essay and should be annotated as such.

A.4 Examples

A.4.1 Example 1

Essay prompt: "Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value."

Text of essay response to prompt:[1]

ICLE FILE #: ICLE-TS-NOUN-0460.1

Annotation of essay stance:
The essay is arguing for *the statement in the prompt.*

A.4.2 Example 2

Essay prompt: "In the words of the old song, 'Money is the root of all evil.' "

Text of essay response to prompt:

ICLE FILE #: ICLE-DB-KVH-0042.3

Annotation of essay stance:
The essay is arguing against *the statement in the prompt.*

A.4.3 Example 3

Essay prompt:
"Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?"

Text of essay response to prompt:

ICLE FILE #: ICLE-RU-MOS-0009.1

Annotation of essay stance:
The essay is NEITHER *arguing* for *nor* against *the statement in the prompt.*

[1] ICLE's usage restrictions do not allow us to print the full text of the essay provided to annotators. Instead, we have provided the file number of the essay.

Appendix B

Supporting argument annotation stance protocol

B.1 Background

This annotation task involves reading sentences taken from student essays and making a decision if a given sentence offers a reason for arguing in agreement or disagreement with a statement. For each sentence, the essay prompt to which the student wrote his/her response will be provided. Your task is to decide if the given sentence is offering a reason for arguing in agreement with the statement in the prompt, is offering a reason for arguing in disagreement with the prompt, or is doing neither.

B.2 Annotation format

Along with the essay sentence, you will be given the original prompt statement to which the student wrote his/her response. Read the sentence and decide if the sentence

- offers a reason for arguing *for* the prompt statement.
- offers a reason for arguing *against* the prompt statement.
- does not offer a reason for arguing either *for* or *against* the prompt statement

B.3 Examples

B.3.1 Example 1

> **Essay prompt**: “Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination.”

Essay sentence: “Another thing that has made people less creative and more passive is the revolution of the computer. ”

Sentence annotation: This sentence offers a reason for arguing in agreement with the prompt statement.

Comment: The writer offers evidence in favor of the prompt’s claim that technology leaves no room for imagination: an alleged link between computer use and diminished creativity.

B.3.2 Example 2

Essay prompt: “Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.”

Essay sentence: “If Marx were alive today he would definitely think that television was the opium of the people.”

Sentence annotation: This sentence does not offer a reason for arguing in either agreement or disagreement with the prompt statement.

Comment: The writer has simply asserted that the prompt statement is correct. He/she has not offered any reason why the statement is correct.

B.3.3 Example 3

Essay prompt: “Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.”

Essay sentence:“Over last five years the higher education has become more practical and allows an active and creative person to find his place in life.”

Sentence annotation: This sentence offers a reason for arguing in disagreement with the prompt statement.

Comment: The writer claims that university education has become more practical–i.e, less theoretical. This is a reason for disagreeing with the prompt’s claim that university education is too theoretical.

B.3.4 Example 4

Essay prompt: “The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.”

Essay sentence: “To them, it is just normal.”

Sentence annotation: This sentence does not offer a reason for arguing in either agreement or disagreement with the prompt statement.

Comment: This sentence is a fragment.

Appendix C

Supporting Argument Summarization Evaluation Protocol

C.1 Background

This annotation task involves reading pairs of sentences taken from student essays and deciding which member of the pair gives a more convincing reason in support of the student's argument stance toward a given prompt statement. For each sentence pair, you will be given two pieces of information:

- The text of the essay prompt to which the student wrote his/her response.
- The argument stance (the writer agrees or disagrees with the prompt) of the essay from which the two sentences were extracted.

Your task is to decide which of the two sentences provides a more convincing reason in support of the provided argument stance. If neither sentence provides a convincing reason, you may write "neither" in the box to indicate that neither sentence provides a convincing reason in support of the provided stance.

C.2 Annotation format

Each pair of sentences has been extracted from a student essay responding to the provided prompt. Students were asked to argue *for* or *against* the prompt. Each sentence pair is presented together with both the prompt to which the student wrote his/her response and the student's overall argument stance (*for* or *against*) toward the statement in that prompt. After reading the essay prompt, examine the two sentences and decide which sentence offers a more convincing reason in support of the overall argument stance. Then, type one of the following three words in the text box provided: *first*, *second*, or NEITHER. Each of these words indicates the following:

- *first* = the first sentence gives a more convincing reason in support of the student's overall stance toward the prompt.

- *second*= the second sentence gives a more convincing reason in support of the student's overall stance toward the prompt.
- NEITHER = neither sentence gives a convincing reason in support of the student's overall stance toward the prompt.

Always evaluate each sentence relative to the overall stance provided. If both arguments support a stance opposite to the overall stance then neither is convincing and you should choose "neither."

C.3 Examples

C.3.1 Example 1

Essay prompt: Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination.

Essay stance: *against*

First sentence: "What I am aiming at is to prove that science, technology and industrialisation regarded as a part of the human progress are nothing but mere results of dreaming and imagination."

Second sentence: "Although I disagree with the statement given above, I think it is quite beautifully formulated."

Sentence annotation: `first`

COMMENT: The original prompt asserts that science, technology, and industrialization have displaced dreaming and imagination. The first sentence provided argues that the opposite is true: science, technology, and industrialization are the products of dreaming and imagination. The second sentence includes the author's stance toward the statement but, unlike the first sentence, does not provide a reason in support of that stance.

C.3.2 Example 2

Essay prompt: All armies should consist entirely of professional soldiers: there is no value in a system of military service.

Essay stance: *for*

First sentence: "There is a possibility of an alternative military service for those who do not want or cannot undergo the military training."

Second sentence: "We'll manage to keep up the same level of defence with considerably smaller amount of people because, unlike young guys who are forced into service against their will, these people would be highly motivated professionals."

Sentence annotation: `second`

COMMENT: In the second sentence, the writer argues that a smaller army comprised of positively motivated volunteer soldiers is a better option than a larger army of conscripts, many of whom have no interest in military service. This is a reason in support of a *for* stance toward the essay prompt. By contrast, the first sentence seems to be a fragment of a larger argument.

C.3.3 Example 3

Essay prompt: In the words of the old song, "Money is the root of all evil."

Essay stance: *for*

First sentence:"You think that the money you get from your parents is not enough, so you try to find a job during the holiday."

Second sentence: "When you are young, you want to study, to go to the university and have a high degree because you know that a good diploma will help you to find easily a good job with a high salary"

Sentence annotation: `neither`

COMMENT: Neither of these sentences provides a reason to support the claim that money is the root of all evil.

Bibliography

Ädel, A. and Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics*, 17(1):3–34.

Anand, P. and Hacquard, V. (2008). Epistemics with attitude. In *Proceedings of SALT*, volume 18, pages 37–54.

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Stroudsburg, PA. Association for Computational Linguistics.

Antinucci, F. and Parisi, D. (1971). On english modal verbs. In *Papers from the 7th Regional Meeting*, volume 7, pages 28–39, Chicago. Chicago Linguistics Society.

Arora, S., Joshi, M., and Rosé, C. P. (2009). Identifying types of claims in online customer reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 37–40, Stroudsburg, PA. Association for Computational Linguistics.

Axelsson, M. W. (2000). USE—the uppsala student english corpus: An instrument for needs analysis. *ICAME journal*, 24:155–157.

Baxendale, P. B. (1958). Machine-made index for technical literature: An experiment. *IBM Journal of Research and Development*, 2(4):354–361.

Besnier, N. (1990). Language and affect. *Annual Review of Anthropology*, 19:419–451.

Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*, volume 23. John Benjamins Publishing, Amsterdam.

Biber, D. (2007). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins, Amsterdam.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Biber, D. and Finegan, E. (1989). Styles of stance in english: Lexical and grammatical marking of evidentiality and affect. *Interdisciplinary Journal for the Study of Discourse*, 9(1):93–124.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education, Harlow, UK.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly, Sebastopol, California.

Branavan, S., Chen, H., Eisenstein, J., and Barzilay, R. (2009). Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, 34(2):569.

Brandow, R., Mitze, K., and Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31(5):675–685.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Cardie, C., Farina, C., Aijaz, A., Rawding, M., and Purpura, S. (2008). A study in rule-specific issue categorization for e-rulemaking. In *Proceedings of the 2008 international conference on Digital government research*, pages 244–253. Digital Government Society of North America.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Kuppevelt, J. and Smith, R., editors, *Current and New Directions in Discourse and Dialogue*, pages 85–112. Springer.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168. ACM.

Chafe, W. and Nichols, J. (1986). *Evidentiality: The Linguistic Coding of Epistemology (Advances in Discourse Processes)*. Ablex Publishing, Norwood, NJ.

Channell, J. (2000). Corpus-based analysis of evaluative lexis. In Hunston, S. and Thompson, G., editors, *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, Oxford.

Chierchia, G. and MacConnell-Ginet, S. (2000). *Meaning and Grammar: An Introduction to Semantics*. MIT press, Cambridge, MA.

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: Volume 2*, pages 590–598, Stroudsburg, PA. Association for Computational Linguistics.

Cilibrasi, R. L. and Vitanyi, P. M. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.

Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the Twelth International Conference on Machine Learning*, pages 115–123.

Conrad, A., Wiebe, J., and Hwa, R. (2012). Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88, Stroudsburg, PA. Association for Computational Linguistics.

Conrad, S. and Biber, D. (2000). Adverbial marking of stance in speech and writing. In Hunston, S. and Thompson, G., editors, *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, Oxford.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons, New York.

Crismore, A. (1989). *Talking with Readers: Metadiscourse as rhetorical act*. Lang, New York.

Crismore, A. and Farnsworth, R. (1989). Mr. Darwin and his readers: Exploring interpersonal metadiscourse as a dimension of ethos. *Rhetoric Review*, 8(1):91–112.

Curran, T. O. J. R. and Koprinska, P. A. I. (2013). An annotated corpus of quoted opinions in news articles. `http://www.tokeefe.org/blog/wp-content/uploads/2013/08/acl13shortopinions.pdf`.

Das, S. and Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, volume 35, page 43.

De Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Stroudsburg, PA. Association for Computational Linguistics.

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 for Computational Linguistics*, pages 984–991, Stroudsburg, PA. Association for Computational Linguistics.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130.

Du Bois, J. (2000). Taking a stance: Constituting the stance differential in dialogic interaction. In *Annual Meeting of the American Anthropological Association*, volume 18, San Francisco.

Du Bois, J. W. (2007). The stance triangle. In Englebretson, R., editor, *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. John Benjamins Publlishing, Amsterdam.

Edmunson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Farkas, R., Vincze, V., Móra, G., Csirik, J., and Szarvas, G. (2010). The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, CoNLL '10: Shared Task, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fellbaum, C. (1998). WordNet: An electronic lexical database.

Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In *Proceedings of the Association for Computational Linguistics*, pages 987–996.

Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003). Background to Framenet. *International Journal of Lexicography*, 16(3):235–250.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 252–259. ACM.

Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics*, 14(3):383–398.

Fraser, B. (2006). Towards a theory of discourse markers. In Fischer, K., editor, *Approaches to Discourse Particles*, pages 189–205. Elsevier, Amsterdam.

Funahashi, T. and Yamana, H. (2010). Reliability verification of search engines hit counts: How to select a reliable hit count for a query. In *Proceedings of the 10th international conference on Current Trends in Web Engineering*, pages 114–125.

Gallie, W. B. (1955). Essentially contested concepts. In *Proceedings of the Aristotelian society*, pages 167–198, London. Harrison and Sons.

Geurts, P. (2005). Bias vs. variance decomposition for regression and classification. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 749–763. Springer.

Givón, T. (1993). *English Grammar: A Function-based Introduction*. John Benjamins Publishing, Amsterdam.

Glaser, A. and Schütze, H. (2012). Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 276–285, Stroudsburg, PA. Association for Computational Linguistics.

Goldberg, A. B. and Zhu, X. (2006). Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52, Stroudsburg, PA. Association for Computational Linguistics.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3):538–546.

Granger, S. and Tyson, S. (1996). Connector usage in the english essay writing of native and non-native EFL speakers of english. *World Englishes*, 15(1):17–27.

Greene, S. C. (2007). *Spin: Lexical Semantics, Transitivity, and the Identification of Implicit Sentiment*. PhD thesis, University of Maryland.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press, Cambridge, MA.

Hacquard, V. (2006). *Aspects of Modality*. PhD thesis, Massachusetts Institute of Technology.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The Weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Halliday, M. and Matthiessen, C. (2004). *An Introduction to Functional Grammar*. Hodder Arnold, London.

Harris, Z. S. (1958). Linguistic transformations for information retrieval. In *Proceedings of the International Conference on Scientific Information*, volume 2, pages 458–71.

Hasan, K. S. and Ng, V. (2013a). Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 816–821, Sofia, Bulgaria. Association for Computational Linguistics.

Hasan, K. S. and Ng, V. (2013b). Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Stroudsburg, PA. Association for Computational Linguistics.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., and de Jong, F. (2011). Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1061–1070. ACM.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Hohmann, H. (2002). Rhetoric and dialectic. In Eemeren, F. V. and Houtlosser, P., editors, *Dialectic and Rhetoric*, pages 41–51. Springer.

Hovy, E. and Lin, C.-Y. (1998). Automated text summarization and the SUMMARIST system. In *Proceedings of the TIPSTER Workshop*, pages 197–214, Baltimore, MD. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Hunston, S. (2007). Using a corpus to investigate stance quantitatively and qualitatively. *Stancetaking in Discourse*, pages 27–48.

Hunston, S. (2010). *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. Routledge, Abingdon.

Hunston, S. and Thompson, G., editors (2000). *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, Oxford.

Hyland, K. (1990). A genre description of the argumentative essay. *RELC Journal*, 21(1):66–78.

Hyland, K. (1998). Exploring corporate rhetoric: Metadiscourse in the CEO's letter. *Journal of Business Communication*, 35(2):224–244.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. Longman, London.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. Continuum International Publishing Group, London.

Hyland, K. and Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, 6(2):183–205.

Hyland, K. and Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, 25(2):156–177.

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 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 296–303.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In Weir, G., Ishikawa, S., and Poonpon, K., editors, *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11. University of Strathclyde Press, Glasgow.

Jakobson, R. (1960). *Linguistics and Poetics*. MIT Press, Cambridge, MA.

Jespersen, O. (1992 [1922]). *The Philosophy of Grammar*. University of Chicago Press, Chicago.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent Twitter sentiment classification. pages 151–160, Stroudsburg, PA. Association for Computational Linguistics.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Jordan, R. R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers*. Cambridge University Press, Cambridge.

Joshi, M. and Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference: Short Papers*, pages 313–316, Stroudsburg, PA. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, Englewood Cliffs, NJ.

Karkkainen, E. (2007). The role of *I guess* in conversational stancetaking. In Englebretson, R., editor, *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. John Benjamins Publlishing, Amsterdam.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, 1(1):3–44.

Keisanen, T. (2007). Stancetaking as an interactional activity: Challenging the prior speaker. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 253–281.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367, Stroudsburg, PA. Association for Computational Linguistics.

Kim, S.-M. and Hovy, E. (2006). Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pages 483–490, Stroudsburg, PA. Association for Computational Linguistics.

Klebanov, B. B., Beigman, E., and Diermeier, D. (2010). Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 253–257, Stroudsburg, PA. Association for Computational Linguistics.

Klein, D. and Manning, C. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Kopple, W. J. V. (1985). Some exploratory discourse on metadiscourse. *College composition and communication*, pages 82–93.

Koshik, I. (2005). *Beyond Rhetorical Questions*. John Benjamins, Amsterdram.

Kwon, N., Shulman, S. W., and Hovy, E. (2006). Multidimensional text analysis for erulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research*, pages 157–166. Digital Government Society of North America.

Labov, W. (1972). *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, Philadelphia.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lee, K. (2013). Korean ESL learners' use of connectors in english academic writing. *English Language Teaching*, 25(2):81–103.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

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 Tenth Conference on Computational Natural Language Learning*, pages 109–116, Stroudsburg, PA. Association for Computational Linguistics.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2010). A PDTB-styled end-to-end discourse parser. Technical report, School of Computing, National University of Singapore.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, H. (2004). Montylingua: An end-to-end natural language processor with common sense.

Liu, H. and Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Lyons, J. (1977). *Semantics*. Cambridge University Press, New York.

Macdonald, C. and Ounis, I. (2006). The TREC Blogs06 collection: Creating and analysing a blog test collection. Technical report, Department of Computer Science, University of Glasgow.

Malouf, R. and Mullen, T. (2008). Taking sides: User classification for informal online political discourse. *Internet Research*, 18(2):177–190.

Mani, I. (2001). *Automatic Summarization*, volume 3. John Benjamins Publishing, Amsterdam/Philadelphia.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Mao, Y. and Lebanon, G. (2006). Isotonic conditional random fields and local sentiment flow. In Schölkopf, B., Platt, J., and Hoffmann, T., editors, *Advances in Neural Information Processing Systems 19*, pages 961–968. MIT Press, Cambridge, MA.

Martin, J. R. (1992). *English Text: System and structure*. John Benjamins Publishing.

Martin, J. R. (2000). Beyond exchange: Appraisal systems in english. In Hunston, S. and Thompson, G., editors, *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, Oxford.

Martin, J. R. and White, P. R. (2005). *The Language of Evaluation*. Palgrave Macmillan, New York.

Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD05, the 9th Pacic-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 301–311.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, volume 752, pages 41–48.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159, Stroudsburg, PA. Association for Computational Linguistics.

McKeown, K. and Radev, D. R. (1995). Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82. ACM.

Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M. R., and Banchs, R. (2010). Opinion mining of Spanish customer comments with non-expert annotations on Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 114–121. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

Mitsakaki, E., Prasad, R., Joshi, A., and Weber, B. (2004). Penn discourse treebank: Annotation tutorial. *Institute for Research in Cognitive Science, University of Pennsylvania*.

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: Volume 2*, pages 599–608. Association for Computational Linguistics.

Moore, R. C. (2009). What do computational linguists need to know about linguistics? In *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, ILCL '09, pages 41–42.

Mosteller, F. and Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, Reading, MA.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 4, pages 412–418, Stroudsburg, PA. Association for Computational Linguistics.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies*, pages 786–794, Stroudsburg, PA. Association for Computational Linguistics.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., and Smeaton, A. F. (2009). Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the*

1st international CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, pages 9–16. ACM.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Stroudsburg, PA. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, pages 79–86, Stroudsburg, PA. Association for Computational Linguistics.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. page 71.

Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

Platt, J. et al. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In Shanahan, J., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–10. Springer, Dordrecht, The Netherlands.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In Kao, A. and S.R.Poteet, editors, *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, pages 9–28.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. `http://snowball.tartarus.org/texts/`.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The PennDiscourse Treebank 2.0. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA).

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. L. (2007). The Penn Discourse Treebank 2.0 Annotation Manual.

Purpura, S., Cardie, C., and Simons, J. (2008). Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 2008 international conference on Digital government research*, pages 234–243. Digital Government Society of North America.

Qin, B., Zhao, Y., Gao, L., and Liu, T. (2008). Recommended or not? Give advice on online products. In *Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08*, volume 4, pages 208–212. IEEE.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Quirk, R., Crystal, D., and Education, P. (1985). *A Comprehensive Grammar of the English Language*. Longman, London.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Stroudsburg, PA. Association for Computational Linguistics.

Reed, C. and Grasso, F. (2001). Computational models of natural language argument. In *International Conference on Computational Science, ICCS 2001*, pages 999–1008. Springer Verlag, San Francisco.

Reed, C., Mochales Palau, R., Rowe, G., and Moens, M.-F. (2008). Language resources for studying argument. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008)*, pages 91–100. European Language Resources Association (ELRA).

Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. (2010). Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, pages 1418–1421.

Renkema, J. (2004). *Introduction to Discourse Studies*. John Benjamins, Amsterdam.

Ross, J. (1969). Auxiliaries as main verbs. In Todd, W., editor, *Studies in Philosophical Linguistics*, pages 77–102. Great Expectations Booksellers and Publishers, Evanston, IL.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ.

Saggion, H., Teufel, S., Radev, D., and Lam, W. (2002). Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, 1(3):261–377.

Sauper, C., Haghighi, A., and Barzilay, R. (2011). Content models with attitude. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1*, HLT '11, pages 350–358, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saurí, R. and Pustejovsky, J. (2009). FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Scheibman, J. (2007). Subjective and intersubjective uses of generalizations in english conversations. In Englebretson, R., editor, *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 111–138. John Benjamins Publlishing, Philadelphia.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622. ACM.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263. Association for Computational Linguistics.

Snyder, B. and Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2007)*, pages 300–307, Stroudsburg, PA. Association for Computational Linguistics.

Somasundaran, S. (2010). *Discourse-level Relations for Opinion Analysis*. PhD thesis, University of Pittsburgh.

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, Stroudsburg, PA. Association for Computational Linguistics.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156. Association for Computational Linguistics.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT press, Cambridge, MA.

Stoyanov, V. and Cardie, C. (2006a). Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 336–344, Stroudsburg, PA. Association for Computational Linguistics.

Stoyanov, V. and Cardie, C. (2006b). Toward opinion summarization: Linking the sources. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 9–14. Association for Computational Linguistics.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1556–1560. ACM.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge.

Swales, J. (2004). *Research Genres: Explorations and Applications*. Cambridge University Press, Cambridge.

Swales, J. and Najjar, H. (1987). The writing of research article introductions. *Written Communication*, 4(2):175–191.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Taboada, M. and Grieve, J. (2004). Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 158–161, Stanford University, CA. AAAI Press.

Taboada, M., Voll, K., and Brooke, J. (2008). Extracting sentiment as a function of discourse structure and topicality. Technical report, Simon Fraser Univeristy School of Computing Science.

Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773.

Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4(3):341–376.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335. Association for Computational Linguistics.

Thompson, G. and Thetela, P. (1995). The sound of one hand clapping: The management of interaction in written discourse. *Text*, 15:103–103.

Thompson, G. and Zhou, J. (2000). Evaluation and organization in text: The structuring role of evaluative disjuncts.

Titov, I. and McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2008)*. ACL.

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, volume 1, page 6.

Toulmin, S. E. (1958 [2004]). *The Uses of Argument*. Cambridge University Press, Cambridge.

Trivedi, R. and Eisenstein, J. (2013). Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 808–813, Stroudsburg, PA. Association for Computational Linguistics.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424, Stroudsburg, PA. Association for Computational Linguistics.

Turney, P. D. 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.

Uszkoreit, H. (2009). Linguistics in computational linguistics: Observations and predictions. In *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, ILCL '09, pages 22–25.

Van Eemeren, F. H. and Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press, Cambridge.

Voll, K. and Taboada, M. (2007). Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence*, pages 337–346, Gold Coast, Australia.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J. (2012a). "That is your evidence?" Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., and King, J. (2012b). A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, pages 812–817. European Language Resources Association (ELRA).

Walton, D. N. (1996). *Argumentation Schemes for Presumptive Reasoning*. Routledge, Mahwah, NJ.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press, New York.

Williams, J. (2007). *Style: Ten Lessons in Clarity and Grace*. Longman, New York.

Wilson, T. and Wiebe, J. (2005). Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 53–60, Stroudsburg, PA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Lingusitics*, 35(3):399–433.

Wilson, T. A. (2008). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes, of Private States*. PhD thesis, University of Pittsburgh.

Witten, I. and Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 25–30, Chicago, IL. AAAI Press.

Wray, A. (2005). *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.

Yang, W. and Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education*, 23(1):31–48.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49. ACM.

Zhou, L., Li, B., Gao, W., Wei, Z., and Wong, K.-F. (2011). Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171, Stroudsburg, PA. Association for Computational Linguistics.

Zirn, C., Niepert, M., Stuckenschmidt, H., and Strube, M. (2011). Fine-grained sentiment analysis with structural features. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2011)*, pages 336–344.