WHAT'S THAT SUPPOSED TO MEAN?
MODELING THE PRAGMATIC MEANING OF UTTERANCES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Marie-Catherine de Marneffe

November 2012

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Christopher Manning, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Daniel Jurafsky

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Beth Levin

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Christopher Potts

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

Abstract

Many strands of natural language processing work, by and large, capture only the literal meaning of sentences. However, in even our most mundane interactions, much of what we communicate is not said explicitly but rather inferred from the context. If I ask a friend to lunch and she replies, *I had a very large breakfast*, I will infer that she does not want go, even though she (perhaps deliberately) avoided saying so directly. This dissertation focuses on building computational models of such pragmatic enrichment. I aim at capturing aspects of *pragmatic meaning*, the kind of information that a reader will reliably extract from an utterance within a discourse.

I investigate three phenomena for which humans readily make inferences. The first study concentrates on interpreting answers to *yes/no* questions which do not straightforwardly convey a 'yes' or 'no' answer. I focus on questions involving scalar modifiers (*Was the movie wonderful? It was worth seeing.*) and numerical answers (*Are your kids little? I have a 10 year-old and a 7 year-old.*). To determine whether the intended answer is *yes* or *no*, we need to evaluate how *worth seeing* relates to *wonderful*, and how *10 and 7 year-old* relate to *little*. Can we automatically learn from real texts what meanings people assign to these modifiers? I exploit the availability of a large amount of text to learn meanings from words and sentences in contexts. I show that we can ground scalar modifier meaning based on large unstructured databases, and that such meanings can drive pragmatic inference.

The second study detects conflicting statements. If an article about a factory says that *100 people were working inside the plant where the police defused the rockets*, whereas a second about the same factory reports that *100 people were injured*, and we understand these statements, we will infer that they are contradictory. I created the first available corpus of contradictions which, departing from the traditional view in formal semantics, I have defined as pieces of text that are extremely unlikely to be considered true simultaneously. I argue that such a definition, rather than a logical notion of contradiction, better fits people's intuitions of what a contradiction is. Through a detailed analysis of such naturally-occurring conflicting statements, I identified linguistic factors which give rise to contradiction. I then used a logistic regression model to learn the best way of weighing these different factors, and put this model to use to predict whether a new set of sentence pairs was contradictory.

The third study targets veridicality – whether events described in a text are viewed as actual, non-actual or uncertain. What do people infer from a sentence such as *At a news conference, Mr. Fournier accused Paribas of planning to pay for the takeover by selling parts of the company*? Is Paribas going to pay for the takeover by selling parts of the company? I show that not only lexical semantic properties but context and world knowledge shape veridicality judgments. Since such judgments are not always categorical, I suggest they should be modeled as distributions. I build and describe a classifier, which balances both lexical and contextual factors and can faithfully model human judgments of veridicality distributions.

Together these studies illustrate how computer systems begin to recover hearers' readings by exploiting probabilistic methods and learning from large amounts of data in context. My dissertation highlights the importance of modeling pragmatic meaning to reach real natural language understanding. Humans rely on context in their everyday use of language. Computer programs must do likewise, and the work presented here shows that it is feasible to automatically capture some aspects of pragmatic meaning.

Acknowledgements

My thanks go first to my committee, without whom this dissertation would not exist. To Chris Manning, my advisor: I am so glad you answered your office phone on Christmas Eve eight years ago and accepted me as a visitor in the NLP group! This was the start of my American journey, and of an amazing time at Stanford. You have been a wonderful advisor. You have shaped my thoughts, always critiqued my work in a constructive way, and helped me gain confidence in myself. You have guided me all the way, put up with my multiple leg injuries, and I know you truly cared about me. For all your guidance, encouragement and support: thank you! To Chris Potts: I am grateful for the many invaluable insights and your very concrete help in my work. Thank you for the fruitful collaboration, and for your constant optimism. To Beth Levin: I am indebted for the help you have always provided me with, intellectually and personally. You have been so generous with your time, reading and commenting on so many of my writings. You caught imprecise or unclear formulations, and helped me communicate better what I meant to say. Your comments were always to the point! To Dan Jurafsky: for your contagious energy, for the very useful feedback on my presentations, for helping me to keep an eye on the big picture, and for teaching me to say 'no'.

This dissertation does not reflect all the work I did during my time at Stanford, and I owe a great debt of gratitude to other faculty members with whom I interacted closely on various projects. To Joan Bresnan: you have always asked the right questions, and interacting with you has been extremely encouraging. To Eve Clark: I have discovered the fascinating aspect of language acquisition thanks to you, and as my children are emerging to language, it is quite captivating to experience "live"

everything I have learned from you. I hope to still be able to do research in that area too. To Meghan Sumner: you opened my eyes to the field of speech perception and experimental design. You also helped me tremendously to find a good balance between personal life, motherhood and work. Thank you for all the wonderful advice, and for believing I could do it all. I am also grateful to Peter Sells who guided me during my first year in the PhD program, and to Tom Wasow for his kindness and advice when I faced important life decisions.

To my mentors in Belgium, Leila du Castillon, Lambert Isebaert, Philippe Delsarte, André Thayse and Cédrick Fairon: you played a significant role in my life and helped me shape my academic path. To Jean-luc Doumont, for teaching me your vision of public speaking, and your invaluable help in preparing my interviews.

To the linguistics and computer science students and postdocs who have been here with me, for your friendship, your advice and your help. Special thanks to Bill McCartney, Jenny Finkel, Nate Chambers, Nola Stephens, Anubha Kothari, Middy Tice, Olga Dmitrieva, Yuan D'Antilio and Spence Green: you have been good listeners when I needed it, and helped me in various ways in my research. And of course to my cohort: Matt Adams, Uriel Cohen Priva, Scott Grimm, Jason Grafmiller and Tyler Schnoebelen, thank you for the great six years together and the warm companionship!

To my Belgian friends who preceded me in the Bay Area or followed a path similar to mine: Maureen Heymans, Cédric Dupont, Virginie Lousse, Lieven Caboor, Sylvie Denuit (the best babysitter ever!), Gaëtanelle Gilquin, Caroline De Wagter, Daniel Wolf, and to Charlene Noll who also counts in this group! Thank you for your tremendous support, you have been an oasis of home and I could not have made it without you. To my American mom, Beverly Pacheco: I am grateful for your generosity, and for making me feel so welcome in your family.

To my parents for their love and upbringing which made me who I am. To my sister, Daphné, and my brothers, Olivier and Guillaume, who I can always count on despite the long distance. To Goéric, my husband, for always believing in me and for pushing us to embark on this American journey eight years ago. And last but not least, to our children, Timothée and Aliénor, whose smiles put everything in perspective and force me to stop and smell the roses along the way.

Contents

Abstract **iv**

Acknowledgements **vi**

1 Introduction **1**

1.1 Different levels of meaning 2

1.2 How humans use language 4

1.3 Early days of natural language understanding 6

1.4 Computational models of pragmatic meaning 9

1.4.1 Underspecification in dialogue: The case of gradable adjectives 12

1.4.2 Detection of conflicting information 13

1.4.3 Assessment of veridicality 14

2 The Stanford dependency representation **16**

3 Learning the meaning of gradable adjectives **24**

3.1 Introduction 24

3.2 Uncertainty in indirect answers 27

3.2.1 Quantifying indirect answers to *yes/no* questions 27

3.2.2 Quantifying uncertain indirect answers 31

3.3 Previous work 33

3.4 Gradable adjectives 36

3.5 Corpus description 37

3.5.1 Types of question-answer pairs 38

3.5.2 Answer assignment 39
3.6 Methods 45
3.6.1 Learning modifier scales and inferring *yes/no* answers 45
3.6.2 Interpreting numerical answers 47
3.7 Evaluation and results 49
3.8 Error analysis 50
3.9 Discussion 57

4 Detection of conflicting information 61
4.1 Introduction 61
4.1.1 Definition of the RTE task 62
4.1.2 Definition of "contradiction detection" 65
4.2 A corpus of conflicting information 66
4.2.1 Annotation guidelines for marking contradictions in the RTE datasets 66
4.2.2 Annotation results 70
4.3 Typology of conflicting information 72
4.4 System description 76
4.4.1 Filtering non-coreferent events 78
4.4.2 Contradiction features 81
4.5 Evaluation and results 85
4.6 Error analysis 87
4.7 Discussion 89

5 Assessment of veridicality 93
5.1 Introduction 93
5.2 Veridicality and modality 97
5.3 Corpus annotation 98
5.3.1 FactBank corpus 99
5.3.2 Annotations from the reader's perspective 101
5.3.3 An alternative scale 104
5.4 Lessons from the new annotations 107

5.4.1 The impact of pragmatic enrichment 107
5.4.2 The uncertainty of pragmatic enrichment 114
5.5 A system for veridicality assessment 120
5.6 Evaluation and results 123
5.7 Error analysis 126
5.8 Discussion 133

6 Conclusion 137

Bibliography 141

List of Tables

3.1 Distribution of answer tags to *yes/no* questions in the Switchboard Dialog Act Corpus. 30

3.2 Types of question-answer pairs, and counts in the corpus. 39

3.3 Mean entropy values and standard deviation obtained in the Mechanical Turk experiment for each question-answer pair category. 42

3.4 Numbers of reviews, words and vocabulary size per rating category in the IMDB review corpus, as well as the average number of words per review. 46

3.5 Summary of precision and recall (%) by type. 51

3.6 Precision, recall, and F1 (%) per response category. 51

3.7 Precision, recall, and F1 (%) per response category for the WordNet-based approach. 51

4.1 Number of contradictions in the first three RTE datasets. 71

4.2 Percentages of contradiction types in the RTE3_dev dataset, the RTE3_test dataset and the real contradiction corpus. 76

4.3 Precision and recall for contradiction detection on RTE3_dev using different filtering strategies. 80

4.4 Precision and recall figures for contradiction detection. 86

4.5 Recall by contradiction type. 87

5.1 FactBank annotation scheme. 99

5.2 Fleiss kappa scores with associated p-values. 103

5.3 An alternative five-tag annotation scheme. 104

5.4 Inter-annotator agreement comparing FactBank annotations with MTurk annotations. 107
5.5 Confusion matrix comparing the FactBank annotations (rows) with our annotations (columns). 108
5.6 Log likelihood values for the training and test data. 123
5.7 Precision, recall and F1 on the subsets of the training data (10-fold cross-validation) and test data where there is majority vote, as well as F1 for the baseline. 125

List of Figures

2.1 Stanford dependencies for the sentence *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.* and graphical representation. 19

3.1 Discourse plan operators for *yes* and *no* answers, as well as for two satellite discourse plan operators. 34

3.2 Design of the Mechanical Turk experiment. 40

3.3 The distribution of some scalar modifiers across the ten rating categories. 48

3.4 Decision procedure for using the word frequencies across rating categories in the review corpus to decide what a given answer conveys. . . 48

3.5 Representation of the scales for the different antonym classes (from Cruse & Togia 1995). 53

3.6 Probabilities of being appropriately described as "little", "young" or "warm", fitted on data retrieved when querying the Web for "little kids", "young kids" and "warm weather". 57

3.7 Correlation between agreement among Turkers and whether the system gets the correct answer. 58

4.1 Example of Question/Answer treatment (from Pasca & Harabagiu 2001). 63

4.2 Dependency graphs of text T *CCN reported that several troops were killed in today's ambush* and hypothesis H *Thirteen soldiers lost their lives in the ambush*, as well as alignment from hypothesis to text. . . 78

5.1 Design of the Mechanical Turk experiment. 103

5.2 Distributions by type. 114
5.3 The subset of '4/6' and '5/5' distributions in which one of the dominant categories was Uu. 116
5.4 The contribution of modal auxiliaries to veridicality judgments. . . . 118

Chapter 1

Introduction

This dissertation investigates how to automatically determine the meaning that humans access when faced with a piece of text. We all read between the lines, and understand far more than just what the string of words in a text means. Consider the following exchange:

(1) A: Does Coupa Café accept credit cards?

B: Suzy said she couldn't use either her Visa or her American Express.

Reading this dialogue, people would probably infer that the answer to A's question is *no*. However to understand from B's answer that Coupa Café does not accept credit cards, several steps are required: we need (i) to realize that B is implicitly answering A's question, (ii) to recognize that not being able to use a Visa or an American Express card is typically incompatible with accepting credit cards, and (iii) to know that Suzy is a reliable source and that one can thus trust what she reports. To understand this simple dialogue and get to what it means, we need to go beyond the strings of words and capture the different inferences that are part of the understanding process. However, so far the field of natural language processing (NLP) has concentrated on successfully analyzing the structure of language rather than the interpretation of meaning in context. Tools such as part-of-speech taggers, parsers, and named-entity recognizers, achieve quite reasonable performance, primarily for English, but also for other languages (e.g., French, German, Arabic, Chinese). More recently, there has been a surge of interest within the computational linguistics community in providing real and robust textual understanding, paralleling what humans do. In particular, there is a desire to avoid domain limitations: any text should be understood, no matter the subject. Tackling such a problem inevitably necessitates dealing with meaning.

But meaning comes in different levels and varieties. By and large, natural language understanding in computational linguistics has focused on the literal meaning of sentences, and systems were limited to a sentence-by-sentence analysis. In this work, I argue that to automatically provide text understanding comparable to what humans achieve, it is essential to go beyond the literal meaning: we need to concentrate on the meaning of language *as it is interpreted in context*, which I call "pragmatic meaning". My dissertation provides some explorations in capturing this type of meaning.

1.1 Different levels of meaning

Two levels of meaning commonly distinguished are the "literal meaning" (or "sentence meaning") and the "utterance meaning" (Levinson 1995). Let's take an often used example to illustrate the difference (adapted from Recanati 2004a; Recanati 2004b): suppose someone asks, at about lunchtime, whether I am hungry, and I reply: *I had a very large breakfast.* The literal meaning of the utterance is the proposition that the speaker had a large breakfast at some point in the past. The literal meaning can be defined as "what is said". Such meaning might also include nonarticulated constituents resulting from saturation (the contextual assignment of values to indexicals and free variables in the logical form of the utterance, Recanati 2004a) or free enrichment (Carston 1988). If we adopt this view, the literal content of the example given here is that the speaker had a large breakfast that particular morning. However, the interlocutor in this exchange will infer that the speaker is not hungry. The interlocutor is able to determine what the utterance implies; he will retrieve an enriched interpretation given the utterance's context.

Levinson (1995) subdivides the "utterance meaning" into the traditional notion of "speaker meaning" and what I will refer to as "pragmatic meaning". Levinson employs the terms "utterance-token meaning" vs. "utterance-type meaning" for these. The "utterance-token meaning" (or speaker meaning) is what a speaker intends to convey in a particular speech context. It refers to the inferences that are drawn in an actual context by specific participants with all their beliefs and intentions. On the other hand, the "utterance-type meaning" (or pragmatic meaning) is a level of *systematic* pragmatic inference based on general expectations about how language is normally used. Words and expressions indeed carry default inferences that are systematically accessed in our everyday use of language. For the same word, context can affect these default inferences. For example, *a handful of grapes* and *a handful of people on the beach at Coney Island* will imply very different estimates: the first in the range of 5 to 20 grapes, and the second hundreds or even thousands of people (Mosteller & Youtz 1990:3).

As Levinson (1995) points out, these three levels of meaning (sentence meaning, utterance-type meaning, and speaker meaning) resonate with the three-way distinction between locutionary, illocutionary and perlocutionary acts proposed by Austin (1962): "the locutionary level corresponds to the level of sentence meaning, the illocutionary to our intermediate layer formed of conventions or habits of use, and the perlocutionary to the level of speaker intentions" (Levinson 1995:94). The classification proposed by Levinson acknowledges the fact that some inferences are systematically drawn, independently of who the speaker and hearer are, or who the writer and reader of a text are. There are patterns of preferred interpretations which arise from the way language is used, and are not linked to the individual participants. I will use the term "pragmatic meaning" to describe such systematic interpretations which we naturally retrieve given some context.

Some researchers actually reject the notion that sentences have a literal meaning. Recanati (2004a) emphasizes the fact that what is said is always dependent on the context, but makes a distinction between what is said and what is implicated. Clark (1996) and Recanati (2004b) point out that some linguistic expressions do not possess a literal meaning as they do not represent anything. Greetings, for instance, frequently only have use-conditions: they are defined in the dictionary as "informal expressions *used* to greet another". There are also sentences in which the literal meaning of some words do not fit, and we can therefore not determine "what is said". For example, Clark argues that in "Diane's approach to life is very *San Francisco*", the literal meaning of *San Francisco* is not appropriate and therefore we need to work out what *San Francisco* means to make sense of the sentence (Clark 1996:144). However, whether we believe or not in literal meaning, and the exact definition we assign to it, is not what matters for my purposes. The important point here is that there is a level of meaning which goes beyond what the strings of words mean, and that this "pragmatic meaning" is a level we access in our everyday understanding of language. For computational language understanding systems to be useful in real world applications, it is essential that they capture the pragmatic meaning of utterances.

1.2 How humans use language

There is lots of evidence that the way humans interpret language is not purely logical. A classic example is the quantifier *some*: whereas its logical interpretation is *at least one, and possibly all*, the ordinary interpretation is *a few, but not all.* Clark (1979) provides the following example in which merchants were asked the question: "Would you mind telling me what time you close tonight?" (pp. 447-448). Six out of the seven people who answered the question (as well as providing the time at which the store closes that day) started their reply by *yes*, which is not a logical answer if only the literal meaning of the question is taken into account. The example demonstrates that people often go beyond the literal meaning of an utterance. Clark argues that the positive answer is triggered by the fact that the merchants *interpret* the question as a request: they work it out as indirectly conveying an intermediate link to a question such as "Will/Can you tell me what time you close tonight?", to which a positive answer is appropriate. In another experiment (Clark & Schober 1992), when merchants were asked on the phone "Do you accept credit cards?", some answered "Uh uh. We're not open anyways." Such an exchange only makes sense when analyzed beneath the surface: it shows that the merchant conceived the caller's plan and inferred that the caller might want to come to the store.

Further, choice of words can induce a change in perspective, which will impact people's interpretation and reaction. Clark & Schober (1992) reports the following example involving the discrepancy between *forbid* and *allow*. When asked the question "Do you think that the United States should forbid public speeches against democracy?", 54% replied yes. When asked the question "Do you think the United States should allow public speeches against democracy?", 75% replied no. Saying yes to the first question is logically equivalent to saying no to the second question. However, the answers received to these two questions were different. Perspective shifts (e.g., presenting data in terms of lives saved versus lives lost) and their effects on the way we reason have been extensively studied by Tversky & Kahneman (1981).

Such examples underscore that more than the surface form matters when processing language. To do so successfully, a human or a machine needs to capture the pragmatic meaning, beyond the literal meaning of the utterance. To take an example closer to standard NLP tasks, what will ultimately define the correct phrase structure parse of a syntactically ambiguous sentence, such as "One morning I shot an elephant in my pajamas" (Groucho Marx)? Where should the prepositional phrase *in my pajamas* be attached: at the level of the verb phrase headed by *shot* or at the level of the noun phrase *an elephant*? The correct parse will be the one reflecting the meaning of the utterance *in context*. In the example given, the ambiguity is actually resolved in the next sentence: "How he got into my pajamas I don't know."

Another aspect of language interpretation is its inherent uncertainty. The way humans interpret language is not purely categorical. There can be some uncertainty in the inferences we draw from utterances. Take the example mentioned above: *a handful of people on the beach at Coney Island*. We are certainly talking about more than 5 people on the beach, but there is some uncertainty as to the number: are we talking here about hundreds or about thousands of people? Meanings are enriched when expressed in context, but this enrichment process might carry some uncertainty. If we want to adequately deal with pragmatic meaning, we need models that capture not only categorical inferences, but also uncertain ones.

1.3 Early days of natural language understanding

In the early days of computational semantics, emphasis was put on providing natural language understanding, and systems that could understand parts of text were successfully developed.

The Chat-80 system (Warren & Pereira 1982) represents the formal semantics tradition. The core idea of that tradition is to represent the meaning of sentences by a logical form. The goal of the Chat-80 system is to understand and answer questions. A question is translated into formal logic, and then transformed into a Prolog query. The query is mapped to a predefined database, and an answer is generated. Words in the lexicon are a set of facts and general rules that define the predicates which correspond to the words. The Chat-80 system was evaluated on a geographical domain, using a relational database containing information about countries, their capitals, their borders, etc. It can accurately, and quickly, answer questions such as *What is the capital of each country bordering the Baltic?* or *How many countries does the Danube flow through?*. However a system such as Chat-80 is limited to a domain linked to a database and for which a lexicon has been defined in order to handle the translation into logic. Also, the translation system only covers a subset of the English language, which on top of the domain specificity creates another limitation. For instance, pronouns are not dealt with. A sentence such as *Which country contains a city bigger than its capital?* is thus beyond the scope of the system. Another drawback of a translation into formal logic is that the process only captures the literal meaning of the sentence composed from the meanings of all its constituents together with the meaning of the constructions in which they occur. In the case of Chat-80, for a question such as *Which ocean borders the United States?*, the presupposition that there is only one answer to the question is ignored, and the system gives all three oceans as the answer.

More recently, the work of Blackburn and Bos still follows the formal semantics tradition (Blackburn & Bos 2005): sentences are translated one-by-one into logical forms to which theorem provers are applied to derive (or not) inferences. When the pragmatic meaning of the sentence aligns with its literal meaning and can be adequately translated into a logical form, such a method yields high precision. However that is rarely the case. For example, in the second Recognizing Textual Entailment Challenge (Bar-Haim *et al.* 2006), where systems are given pairs of sentences and have to decide whether the second sentence can be inferred from the first one, only 3.6% of the test data could be handled by this method, and, even then, only with 76% accuracy (Bos & Markert 2006). Such figures emphasize how much the domain specificity of the method is an issue. It requires tremendous hand-coded efforts to broaden the coverage of the translation system, and to add the knowledge necessary for inferences to go through.

The Schankian tradition adopts another approach: it suggests that understanding text requires characteristic knowledge coded in structured representations of events, with their participants and their causal relationships, called *scripts*. For instance, Schank & Abelson (1977) propose the "restaurant script" to understand texts about restaurants. The restaurant script is a stereotypical representation of a restaurant: it includes the events that constitute a visit to a restaurant (entering the restaurant, choosing a table, sitting down, asking for menus, ordering, etc.), its participants (customer, waiter, tables, cook, etc.), as well as the preconditions and results of the typical actions that occur. Scripts were central to research in the 1970s and 1980s and used in tasks such as summarization, coreference resolution and question answering. The Schankian tradition has the strength that it is much more directed to modeling pragmatic meaning. Scripts target contextual understanding and interpretation beyond literal meaning: they encode plausible inferences rather than only strict logical deductions. However scripts do not capture the uncertainty inherent to interpretation, and the inferences remain categorical. Another big limitation of this approach is that it can only handle situations which fully adhere to the scripts. It is thus extremely specific and limited to domains for which scripts have been defined. It also requires a tremendous amount of hand-coding.

Peter Norvig's dissertation proposes a unique algorithm which uses common-sense knowledge to make different kinds of inferences from texts (Norvig 1986). Instead of having to create a new script for a new situation, the idea is to populate a knowledge base with the necessary information to handle the new situation. As Norvig acknowledges, if his approach is less brittle than scripts when contexts slightly change, it shifts the work towards the knowledge database. A new script does not need to be defined to handle a new situation, but the knowledge required to reason about it needs to be present in the database. Norvig's approach, like scripts, thus has restricted coverage in terms of domain and still needs to code knowledge by hand. The knowledge base is a network of concepts where the links represent connections between the concepts. The algorithm makes use of the knowledge base to draw several kinds of inferences from texts. One major class is "referential inference", in other words, the system does coreference resolution. It also deals with "elaboration inference". For example, in a sentence such as *hoping to catch a few fish from the sea, which they could sell*, the algorithm will infer the event *having* as a result of the "catching" event, and a precondition of the "selling". Another category of inference is the "view application inference": in *The Red Sox killed the Yankees*, the system will interpret *kill* as a *defeat* and not as an instance of actual murder. "Concretion inferences" can also be drawn: in *John cut the grass*, *cut* is interpreted as a *lawn-cutting*, i.e. mowing, event, and not just as any kind of cutting, such as with scissors or a knife for instance. The same process occurs for disambiguating modifiers, such as in (2) where the modifier introduced by the preposition *with* is ambiguous. The system will be able to disambiguate between an accompanier, an instrument, a manner or a default "along with" modifier (in this specific case, the system will know that *pesto* is a sauce and will choose this more specific interpretation).

(2) a. John ate spaghetti with Frank.

b. John ate spaghetti with a fork.

c. John ate spaghetti with gusto.

d. John ate spaghetti with pesto.

Norvig's system captures a broad set of inferences, but still fails to recognize that inferences can be uncertain, and only provides categorical inferences.

Early approaches to natural language understanding fall short in adequately capturing pragmatic meaning. To sum up, there are five major drawbacks that need to be dealt with to provide real text understanding: (i) early approaches to natural language understanding are strongly domain limited, (ii) they require hand-coded effort, (iii) they are primarily concerned with the literal meaning of sentences, (iv) they are often based on strictly logical inference, rather than on common-sense reasoning, and (v) they do not allow uncertain inference. The approach I am taking will aim at a broad coverage, no hand coding, and the pragmatic level of meaning. It will also focus on the uncertainty that is inherently present when targeting the level of pragmatic meaning.

1.4 Computational models of pragmatic meaning

To approximate human understanding, it is essential for NLP systems to access the level of pragmatic meaning, which targets the systematic inferences that readers/hearers commonly draw when faced with language. Recent work in sentiment analysis is, in a sense, working at the level of pragmatic meaning (i.a., Pang *et al.* 2002; Pang & Lee 2004; Pang & Lee 2008; Fahrni & Klenner 2008; Choi *et al.* 2009). The goal of sentiment analysis is to identify the viewpoints underlying a text span. Sentiment analysis refers to any computational treatment of opinion, sentiment, and subjectivity in text. Sentiment analysis often focuses on determining an overall sentiment, e.g., determining whether a review is positive or negative, by examining whether words bear a positive or negative connotation. Sentiment analysis also tries to determine the different opinions people attribute to something. Even though the techniques developed to retrieve the polarity of words do not rely on semantic analysis (they mainly spot key words and key phrases), the work done in sentiment analysis emphasizes that the polarity is often highly dependent on *context*. For example, the adjective *unpredictable* carries a positive polarity in an *unpredictable plot* in the movie domain, but has a negative connotation in an *unpredictable boss*. There are words whose polarity is domain-independent, such as the positive adjectives *excellent*, *perfect*, *wonderful* (though figurative language can produce counter-examples). But in general, the polarity varies across domains. It can even vary within a specific domain: in the domain of food, *cold beer* will be viewed as positive whereas *cold pizza* will not. Sentiment analysis provides a simple, but clear case of how meanings are systematically enriched when expressed in context: the connotation given to a word can shift from positive to negative according to the context. The need to analyze language *as it is used*, grounded in the context of the utterance, is thus essential to achieving accurate sentiment analysis.

The central aim of this dissertation is to propose computational models of pragmatic meaning which rely on semantic analysis. I show through annotations tapping into pragmatic meaning that the interpretations people have are systematic enough to be the focus of computational work on textual understanding. Yet, the annotations also reveal that people are not always sure what the interpretation is. Probabilistic models are inherently well-suited to dealing with uncertainty, and therefore, throughout my work, I adopt models that are probabilistic to adequately capture the uncertainty intrinsic to pragmatic enrichment. Modeling uncertainty with probabilities is in line with a body of work in psychology which shows that uncertainty is represented probabilistically in human cognition (i.a., Reyna 1981; Mosteller & Youtz 1990; Lassiter 2011).

If we want to analyze language *as it is used*, it is essential to work with corpus data which occurs naturally, instead of with constructed examples which omit all sorts of nuances. My work is thus grounded in naturally-occurring data. I also make use of crowdsourcing techniques, which are ideal for exploring pragmatic meaning, where intuitions about language are the primary data. Crowdsourcing techniques allow us to tap into people's intuitions and to quickly obtain a large number of annotations from a wide population.

Modern work on language understanding offers statistical techniques to learn diverse knowledge automatically from free text. Further, in this era of rapidly expanding user-generated web content, we now have access to large amounts of situated language. I therefore exploit such statistical techniques paired with the availability of large amounts of text to learn pragmatic meanings from words and sentences appearing in real contexts, instead of requiring hand coding. My work shows that this is a successful approach to building models that capture word meaning.

Specifically I will concentrate on three kinds of inferences people commonly draw in their everyday use of language: (i) inferences arising from underspecification in dialogue, (ii) inferences that are drawn between pieces of text, and (iii) inferences made about the status of an event. The first type of inference targets the retrieval of implicit information: information is indeed often left underspecified in language, but readers/hearers retrieve what is meant. For instance, even though B's answer in (3), repeated from (1), does not explicitly contain a *yes* or a *no*, people will readily infer whether B's reply conveys a *yes* or a *no* answer.

(3) A: Does Coupa Café accept credit cards?

B: Suzy said she couldn't use either her Visa or her American Express.

The second type of inference targets the relationships people draw between pieces of text. Sentences or passages rarely stand alone, but are part of a text: people thus make connections between them. In conversation, for example, people will try to find a coherent link between two utterances. A standard example to illustrate this point comes from Grice (1975): if A utters *I am out of petrol* and his interlocutor replies *There is a garage round the corner*, A will expect that answer to be relevant to his problem; he will expect that his interlocutor thinks the garage is open and sells petrol. This search for coherence is not limited to dialogue. When reading different articles, we ask ourselves how they relate to each other: do they convey similar or conflicting information, or are they simply unrelated? The third type of inference concerns the status of an event: when an event is described, people assess whether or not it corresponds to a fact in the real world. For example, what do people infer from a sentence such as *At a news conference, Mr. Fournier accused Paribas of planning to pay for the takeover by selling parts of the company*? Is Paribas going to pay for the takeover by selling parts of the company? These three types of inference capture ways people enrich the meaning of utterances in relation to their context. To achieve text understanding (whether we are concerned with written text or speech), NLP needs to be able to model such meanings—the ones that correspond to enriched utterance interpretations. In the remainder of this chapter, I introduce the three case studies that compose my dissertation.

1.4.1 Underspecification in dialogue: The case of gradable adjectives

The first case study focuses on information that is left underspecified in the literal meaning of what is said, but that we nonetheless infer. In particular, I focus on interpreting those answers to *yes/no* questions which do not straightforwardly state *yes* or *no*. When a speaker does not explicitly answer such a question with *yes* or *no*, participants in the conversation will try to infer what the intended reply is. I concentrate on questions which contain a gradable modifier and whose answers include another gradable modifier (4) or with a numerical expression (5).

(4) A: Was the movie wonderful?

B: It was worth seeing.

(5) A: Are your kids little?

B: I have a 10 year-old and a 7 year-old.

To determine whether the intended answer is ‘yes’ or ‘no’ in the above examples, we need to evaluate how *worth seeing* relates to *wonderful*, and to determine whether a *10 year-old* and a *7 year-old* are considered to be *little*. The goal of this case study is to show that we can automatically learn from real texts the scalar orderings people assign to these terms, and infer the extent to which a given answer conveys ‘yes’ or ‘no’. The chapter presenting this case study is based on two previously published papers (de Marneffe *et al.* 2009; de Marneffe *et al.* 2010).

Since in some of these question-answer pairs, the intended answer cannot be categorically determined, as in (6), I develop probabilistic models to learn scalar orderings based on data collected from the Web. To evaluate the methods I develop, I first collect naturally-occurring examples of question-answer pairs involving gradable modifiers. I then use response distributions from crowdsourcing workers to assess the degree to which each answer in the corpus conveys ‘yes’ or ‘no’. The experimental results closely match the workers’ response data, demonstrating that meanings can be successfully learned from Web data and that such meanings can drive pragmatic inference.

(6) A: Was it acceptable?

B: It was unprecedented.

1.4.2 Detection of conflicting information

The second case study targets inferences that are drawn between pieces of text. We constantly make links between different chunks of text, assessing how one relates to another. For instance, despite the lack of word overlap between the two sentences in (7), we infer that they describe the same event. However to identify the connection between the sentences, we need to go beyond the literal meaning and use enriched meaning. We need to know that offering a vacation stay to a senator is bribery.

(7) (a) In return for political favors, Jack Abramoff offered the senator a vacation stay.

(b) Lobbyist attempts to bribe a U.S. legislator.

In contrast, if an article about a factory says that *100 people were working inside the plant where police specializing in explosives defused the rockets*, whereas a second about the same factory reports that *100 people were injured*, we infer that they are contradictory. Again such understanding requires pragmatic meaning: defused rockets cannot go off, and thus cannot injure anyone. I concentrate on the automatic detection of contradictions between text passages, and build a system to do so. The chapter describing this second case study is largely based on the material presented in de Marneffe *et al.* (2008) and de Marneffe *et al.* (2011b).

The main contribution of this chapter is a definition of contradiction suitable for NLP. Departing from the traditional view in formal semantics in which two sentences are contradictory if there is no possible world in which both sentences are true, I define contradictions as pieces of text that are extremely unlikely to be considered true simultaneously. I argue that such a definition, rather than a logical notion of contradiction, latches on to pragmatic meaning and therefore better fits people's intuitions of what a contradiction is.

I have created the first available corpus of naturally-occurring contradictions. Through a detailed analysis of the conflicting statements in the corpus, I identified patterns which give rise to contradiction and propose a typology of contradictions. I also give the first detailed error analysis for the contradiction detection task, breaking down performance by contradiction type.

1.4.3 Assessment of veridicality

In the third case study, I examine veridicality – whether events mentioned in a text are viewed by readers as actually happening, not happening or whether it is unclear if the events happened. What do people infer from a sentence such as *FBI agents alleged in court documents today that Zazi had admitted receiving weapons and explosives training from al Qaeda operatives*? Do people take it as true that Zazi received weapons and explosives training? The source of the allegation may affect our judgment about whether such events occurred. We might react differently if the sentence states that the source is an anonymous tip, rather than FBI agents. Yet, there is a long tradition in formal semantics of simply assuming that the complement of *allege* is non-veridical. Such theories which assume a lexical item uniquely assigns a veridicality value to its complement neglect the sort of pragmatic enrichment that is pervasive in human communication.

This chapter extends the work presented in de Marneffe *et al.* (2011a) and de Marneffe *et al.* (2012). To understand what contributes to pragmatic enrichment in the context of veridicality, I used crowdsourcing techniques to gather people's veridicality judgments of events. I show that not only lexical semantic properties of words but also context and world knowledge shape their judgments. Since judgments for a given event are not always unanimous, I model them as distributions over the possible veridicality values, instead of assigning them a single value. I build a classifier to automatically assign event veridicality distributions. The classifier balances both lexical and contextual factors and can faithfully model human judgments of veridicality.

This work argues that to accurately understand text, it is crucial to shift the focus from literal meaning to the level of pragmatic meaning. An important challenge for researchers in NLP is to learn not only basic linguistic meanings but also how those meanings are systematically enriched when expressed in context. My dissertation explores how to automatically model such pragmatic meaning. The approach I take shows how to successfully combine probabilistic models with linguistic analysis, exploiting the modern data-rich world we now have access to.

Chapter 2

Background: The Stanford dependency representation

Before moving on to the core of the dissertation, I briefly introduce a linguistic representation that is either mentioned or heavily relied on in each of the following chapters. The linguistic representation I am working with is the Stanford typed dependency representation (de Marneffe *et al.* 2006).

The scope of analysis of a dependency representation is the sentence. Each word in the sentence is related to other words in the sentence which depend on it. A typed dependency representation labels the dependencies between individual words of the sentence with grammatical relations (e.g., *subject*, *indirect object*). Dependency representations have been the dominant form of linguistic representation throughout history. Even though modern work on dependency representations often links to the work of Lucien Tesnière (Tesnière 1959), such dependency representations exist since the earliest recorded grammars (e.g., Panini's grammar ∼6th century BCE) and were used by the first millennium Arabic grammarians. In NLP, some of the earliest kind of parsers made use of a dependency representation: David Hays, one of the founder of computational linguistics, built a dependency parser in the 1960s (Hays & Ziehe 1960; Hays 1964). The constituency representation, which is nowadays the dominant representation in linguistics as well as a prominent representation in NLP, is actually a recent invention. It started with the immediate constituent (IC) analysis in the

1930s (Bloomfield 1933; Wells 1947) and developed to its present place of dominance through Chomsky's work on phrase structure grammars (Chomsky 1957).

Recently, there has been a big swing back to dependency representation in a wide range of NLP tasks. Many applications in NLP are primarily interested in predicates and their argument structure, and benefit particularly from having access to a shallow meaning representation. Constituency representations are otiose for meaning recovery (predicate-argument structure are not readily available from phrase structure parses), whereas dependency representations provide an easy way to describe the scope of arguments. Dependency representations also abstract away from various phenomena which do not impact shallow meaning, such as word order, and the presence of some modifiers. For example, the sentences *"I feel like a little kid," says a gleeful Alex de Castro* and *A gleeful Alex de Castro says: "I feel like a little kid"* will have very different constituency representations but an identical dependency representation. For various tasks, such as machine translation, question answering and textual entailment, it is thus much more natural to use a dependency representation than a constituency representation. The advantages that dependency representations have can benefit not only NLP but also linguistic semantics.

The Stanford dependency scheme was designed to provide a simple description of the grammatical relationships in a sentence that could easily be understood and effectively used by people without linguistic expertise who wanted to extract textual relations (de Marneffe & Manning 2008). Unlike many linguistic formalisms, excessive detail is viewed as a defect: information that users do not understand or wish to process detracts from usability. The Stanford dependency scheme tries to provide semantically useful relations, while striking the right balance between human readability and interpretation. The set of relations cannot be too impoverished. A coarse set of relations, such as the one proposed by the CoNLL dependency scheme (Buchholz & Marsi 2006), another widely used dependency representation, improves human readability, but impedes interpretation. Basically the CoNLL dependency scheme only makes a distinction between verb and noun modifiers. However, to obtain a shallow meaning representation useful in practical applications, more fine-grained distinctions are necessary. NP-internal relations are an inherent part of corpus texts and are critical in real world applications. For example, in a noun phrase such as *A gleeful Alex de Castro, a car salesman, who has stopped by a workout of the Suns to slip six Campaneris cards to the Great Man Himself to be autographed*, it is important to be able to differentiate the various modifiers of *Castro*: the adjectival modifier *gleeful*, the apposition *salesman*, the noun compound modifier *Alex* and the relative clause. Such distinctions are therefore present in the Stanford dependency representation.

Figure 2.1 gives an example of the Stanford dependencies for the sentence *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.* All information is represented as binary relations: a dependency is a grammatical relation that holds between a governor and a dependent. The numbers appended to the words indicate their position in the sentence. These dependencies map straightforwardly onto a directed graph representation, in which words in the sentence are nodes in the graph and grammatical relations are edge labels.

The labels of the Stanford dependency representation bears a strong intellectual debt to the framework of Relational Grammar (RG, Perlmutter 1983) as well as Lexical-Functional Grammar (LFG, Bresnan 2001), and, more directly, it owes a debt to both the sets of grammatical relations and the naming defined in two representations that follow an LFG style: the Grammatical Relations (GR) scheme (Carroll *et al.* 1999) and the Palo Alto Research Center (PARC) scheme (King *et al.* 2003) which were used as a starting point for developing the Stanford dependencies. But where the Stanford dependency scheme deviates from GR, PARC, and its RG and LFG roots is that it has been designed to be a practical model of sentence representation, particularly in the context of relation extraction tasks.

The Stanford dependency representation makes available several options, suited to different goals: in one, every word of the original sentence is present as a node with relations between it and other nodes, whereas in the latter, certain words are "collapsed" out of the representation, making such changes as turning prepositions or conjunctions into relations (as can be seen in Figure 2.1). The former is useful when a close parallelism to the source text words must be maintained, such as when used as a representation for direct dependency parsing (Kübler *et al.* 2009), whereas the

nsubj(makes-8, Bell-1)
nsubj(distributes-10, Bell-1)
partmod(Bell-1, based-3)
nn(Angeles-6, Los-5)
prep_in(based-3, Angeles-6)
conj_and(makes-8, distributes-10)
amod(products-16, electronic-11)
conj_and(electronic-11, computer-13)
amod(products-16, computer-13)
conj_and(electronic-11, building-15)
amod(products-16, building-15)
dobj(makes-8, products-16)
dobj(distributes-10, products-16)

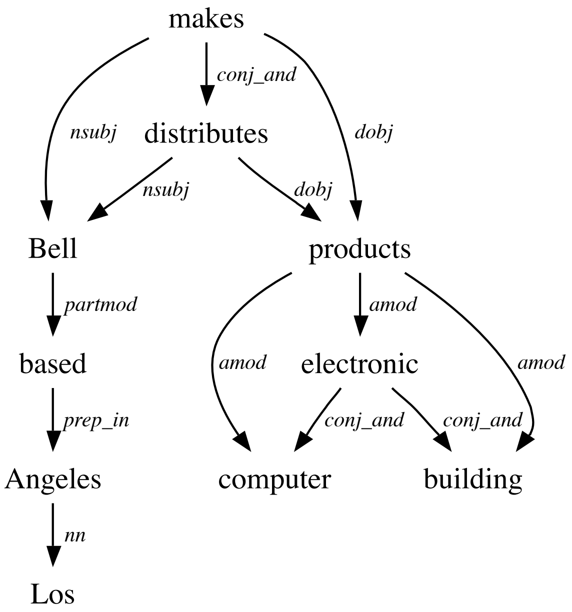


Figure 2.1: Stanford dependencies for the sentence *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.* and graphical representation.

latter is intended to be more useful for relation extraction and language understanding tasks. I am using the latter option, which aims to produce a representation closer to the semantics of the sentence. This collapsed representation adheres to the following design principles:

1. Every relation is represented uniformly as binary (between two sentence words).
2. Relations should be semantically contentful and useful to applications.
3. Where possible, relations should use notions from traditional grammar for easier comprehension by users.
4. Underspecified relations should be available to deal with the complexities of real text, including ungrammaticality and domain-specific usages.
5. Where possible, relations should be between content words, not indirectly mediated via function words.
6. The representation should be minimal rather than overwhelming with unnecessary linguistic details.

In order to automatically assign dependency representation to sentences, I have designed a tool, described in de Marneffe *et al.* (2006), which provides for the rapid extraction of the grammatical relations (i.e., subject, object) from phrase structure parses. The tool is integrated with the Stanford parser (Klein & Manning 2003).[1] Structural configurations over phrase structure parses are used to define the grammatical relations: the semantic head of each constituent of the parse is identified, using rules akin to the Collins head rules (Collins 1999), but modified to retrieve the semantic head of the constituent rather than the syntactic head. As mentioned, content words are chosen as heads, and all the other words in the constituent depend on this head. To retrieve adequate heads from a semantic point of view, heuristics are used to inject more structure when the Penn Treebank gives only flat constituents, as is often the case for conjuncts, e.g., (NP *the new phone book and tour guide*), and QP constituents, e.g., (QP *more than 300*). Then, to retrieve instances of each grammatical relation, I have created patterns over the phrase structure parse tree using

[1] http://nlp.stanford.edu/software/lex-parser.shtml

the tree-expression syntax defined by tregex (Levy & Andrew 2006). Conceptually, each pattern is matched against every tree node, and the matching pattern with the most specific grammatical relation is taken as the type of the dependency.

The results of the automatic relation extraction are not always accurate. For instance, in the sentence *Behind their perimeter walls lie freshly laundered flowers, verdant grass still sparkling from the last shower, yew hedges in an ecstasy of precision clipping*, the system will erroneously retrieve apposition relations between *flowers* and *grass*, as well as between *flowers* and *hedges* whereas these should be *conj_and* relations, indicating a conjunct relation between two elements connected by the coordinating conjunction *and*. The system fails when there is no overt maker of conjunction. In the example sentence from Figure 2.1, the system will actually fail to retrieve the direct object relation between *distributes* and *products*. Another limitation of the tool is the treatment of long-distance dependencies, such as *wh*-movement and control/raising: the system cannot handle long-distance dependencies that cross clauses. In a sentence like *What does he think?*, the system will correctly find that *what* is a direct object of *think*:

dobj(think-4, What-1)
aux(think-4, does-2)
nsubj(think-4, he-3)

However in a sentence such as *Who the hell does he think he's kidding?*, the automatic extraction process will fail to find that *who* is the direct object of *kidding*. Here, it is vital to distinguish between the appropriate dependency representation of a sentence and the dependencies that the conversion tool retrieves. Long-distance dependencies are not absent from the formalism, but the tool does not accurately deal with them.[2]

Yet the Stanford dependency representation and the tool have been successfully used by many researchers in different domains. Perhaps the most striking example of its success comes from the biomedical world. Pyysalo *et al.* (2007) developed a version of the BioInfer corpus annotated with the Stanford dependency representation, which

[2] As possible future work, I am considering using a tool such as Levy & Manning's (2004) to correctly determine long distance dependencies, as input to the current dependency conversion system. This would presumably be effective, but would make the conversion process much slower.

is the main source of gold-standard Stanford dependency data currently available. Shortly after its creation, the conversion tool was used for relation extraction of biomedical entities (Erkan *et al.* 2007; Urbain *et al.* 2007; Fundel *et al.* 2007; Garten 2010; Björne & Salakoski 2011; Pyysalo *et al.* 2011; Landeghem *et al.* 2012) and gradually became the de facto standard in biomedical relation extraction: in the BioNLP 2009 Shared Task, many of the leading teams built their relation extraction systems over the Stanford dependency representation (Kim *et al.* 2009), and in the BioNLP 2011 shared task, every team used it (Kim *et al.* 2011).

The Stanford dependency scheme has been used to evaluate parsers: first in the biological information extraction domain (Clegg & Shepherd 2007; Pyysalo *et al.* 2007), and more recently in the 2012 shared task on parsing the web (Petrov & McDonald 2012). It is a common representation for extracting opinions, sentiment, and relations (Kessler 2008; Haghighi & Klein 2010; Hassan *et al.* 2010; Joshi *et al.* 2010; Wu & Weld 2010; Zouaq *et al.* 2010), as well as specific information (such as event, time or dialogue acts) (Chambers 2011; McClosky & Manning 2012; Klüwer *et al.* 2010). The tool has been consistently used by several groups in the different challenges targeting textual entailment (Malakasiotis 2009; Mehdad *et al.* 2009; Shivhare *et al.* 2010; Glinos 2010; Kouylekov *et al.* 2010; Pakray *et al.* 2011). It is also used for a variety of other tasks, such as coreference resolution, disagreement detection and word sense induction (Chen & Eugenio 2012; Abbott *et al.* 2011; Lau *et al.* 2012), as well as part of the preprocessing for machine translation systems by several groups including Google (Xu *et al.* 2009; Genzel 2010; Sing & Bandyopadhyay 2010). The Stanford dependency representation has also served as a model for developing dependency schemes in other languages. Recently schemes based on the Stanford dependency representation have been proposed for Finnish (Haverinen *et al.* 2010a; Haverinen *et al.* 2010b), Thai (Potisuk 2010), Persian (Seraji *et al.* 2012), and French (El Maarouf & Villaneau 2012).

Recent work in NLP has clearly endorsed the advantages that dependency representations have over constituency representations, and the Stanford dependency scheme in particular has received sustained acclaim for its usability. The next chapters describe my work in automatically capturing pragmatic meaning, the central goal of my dissertation. But throughout this dissertation, the data representation I am depending on is the Stanford dependency scheme. Such a representation is highly advantageous for pursuing pragmatic meaning.

Chapter 3

Learning the meaning of gradable adjectives

3.1 Introduction

In chapter 1, I emphasized that a central issue in capturing pragmatic meaning is that much of the information that humans retrieve when using language is actually left out or underspecified in the literal meaning of what is said. Here, I focus on a particular aspect of this issue, looking at information that is implicit in a conversation, but that participants nonetheless infer. Such underspecification is pervasive in conversations. One example is the possessive construction in English. If I say "John's book is on the table", I do not explicitly specify whether John is the author of the book or its possessor. Both interpretations can exist, but often the context in which the sentence is uttered will suffice to resolve what is meant, and participants in the conversation will easily construct the correct meaning. In this chapter, I investigate another instance of implicit information: indirect answers to polar (*yes/no*) questions, that is, answers which do not explicitly contain a *yes* or *no*, but rather convey information that can be used to infer such an answer with some degree of certainty. The lunchtime example given in the introduction, repeated here in (8), is an instance of an indirect answer to a *yes/no* question:

(8) A: Do you want to go for lunch now?

B: I had a very large breakfast.

In this dialogue, B does not explicitly reply with *yes* or *no*, but A will have to decide whether B's answer means 'yes' or 'no'. Identifying such inferences is fundamental to dealing with conversational implicatures in generation and interpretation (see Hirschberg 1985). In some cases, determining the intended answer is straightforward as in (9) where a 'yes' answer to the question is certainly intended, but in others such as (10), what to infer from the answer is unclear. For instance, Hockey *et al.* (1997) find that, in the Edinburgh map task corpus (Anderson *et al.* 1991), 27% of answers to polar questions do not contain an explicit *yes* or *no* word, and that 44% of these fail to clearly convey a 'yes' or 'no' answer.

(9) A: Was it bad?

B: It was terrible.

(10) A: Was it good?

B: It was provocative.

I first conduct a corpus study to quantify the pervasiveness of indirect answers to *yes/no* questions, as well as the degree of uncertainty in the answers. I then focus on the interpretation of a specific case of question-answer pairs: ones in which the main predication involves a gradable modifier (e.g., *good*, *unusual*, *little*) and the answer either involves another gradable modifier as in (11) or a numerical expression (e.g., *seven years old*, *twenty acres of land*) as in (12). Interpreting such question-answer pairs requires dealing with modifier meanings, specifically, learning context-dependent scales of expressions (Horn 1972; Fauconnier 1975) that determine how, and to what extent, the answer as a whole resolves the issue raised by the question.

(11) [sw00utt/sw_0069_3144.utt][1]

[1] Most of the examples I use in this dissertation come from corpora. The information between brackets is the corpus reference.

A: Was that [the movie] good?

B: Hysterical. We laughed so hard.

(12) [sw05utt/sw_0556_3317.utt]

A: Have you been living there very long?

B: I'm in here right now about twelve years.

I propose two methods for learning the knowledge necessary for interpreting indirect answers to questions involving gradable adjectives, depending on the type of predications in the question and the answer. The first technique deals with pairs of modifiers: I hypothesized that online, informal review corpora in which people's comments have associated ratings would provide a general-purpose database for mining relative orders of modifiers along the relevant scale. I thus use a large collection of online product reviews to learn orderings between adjectives (e.g., *good* < *excellent*), and employ this scalar relationship to infer a *yes/no* answer (subject to negation and other qualifiers). The second strategy targets numerical answers. Since it is unclear what kind of corpora would contain the relevant information, I turn to the Web in general: I use distributional information retrieved via Web searches to assess whether the numerical measure counts as a positive or negative instance of the adjective in question. For example, I query the Web to retrieve instances of children considered little (1 year-old, 5 year-old, etc.) as well as instances of children who are not considered little (18 year-old). Using these positive and negative instances, I build a logistic regression model, which can then predict for any age whether a child of that age will be considered little or not. Both techniques exploit the same approach: using large amounts of text (the Web) to learn meanings that can drive pragmatic inference in dialogue. To evaluate the methods I built a small corpus of question-answer pairs involving gradable modifiers, and through crowdsourcing, obtained judgments of whether the indirect answer conveys a 'yes' or 'no'. I then compared these judgments to the ones obtained automatically. Overall the methods yield 71% accuracy, demonstrating that meaning can be grounded by using large unstructured databases.

3.2 Uncertainty in indirect answers

Previous corpus studies have looked at how pervasive indirect answers to *yes/no* questions are in dialogue. Stenström (1984) analyzed 25 face-to-face and telephone conversations and found that 13% of answers to polar questions do not contain an explicit *yes* or *no* term. In a task dialogue, Hockey *et al.* (1997) found that 27% of the responses to polar questions were indirect. This higher percentage might reflect the genre difference in the corpora used: task dialogue vs. casual conversations. Hockey *et al.* (1997) analyzed eight dialogues averaging seven minutes each from the Edinburgh map task corpus (Anderson *et al.* 1991). The dialogues were carried out in an experimental setting, in which each participant has a schematic map in front of them, not visible to the other. One participant has a route drawn on her map, the second does not. The task is for the participant without the route to draw one on the basis of discussion with the participant with the route. Hockey *et al.* (1997) further analyzed the types of answers, coding for whether the answer as a whole conveyed a 'yes' or a 'no', or a meaning that does not appear to commit to either (non-committal answers). They marked 44% of the indirect answers as non-committal. To assess these figures for more casual conversations, I conducted a corpus study using the Switchboard Dialog Act Corpus (Jurafsky *et al.* 1997).

3.2.1 Quantifying indirect answers to *yes/no* questions

The Switchboard Dialog Act Corpus has been annotated for approximately 60 basic dialog acts (question, answer, backchannel, agreement, disagreement, apology, etc.), clustered into 42 tags. It is a subset of 1155 5-minute conversations from the Switchboard database which features two-sided telephone conversations among speakers from all areas of the United States (Godfrey *et al.* 1992). Topics for the conversations were provided (e.g., death penalty, drug use). The dialog act tags ease the search for *yes/no* questions. I am concerned only with *direct yes/no* questions, and not with indirect ones such as "May I remind you to take out the garbage?" or "Can you reach the salt?" which are conventional ways to indirectly request someone to do a particular act, but do not per se require a 'yes' or 'no' answer (Clark 1979;

Perrault & Allen 1980). From 200 conversations of the Dialog Act Corpus, I extracted direct *yes/no* questions (tagged “qy”) and their answers, but discarded disjunctive questions, such as (13), since these do not necessarily call for a ‘yes’ or ‘no’ response. I also discarded tag questions, such as (14), (15) or (16), because these just ask for confirmation. I also did not take into account questions that were lost in the dialogue, nor questions that did not really require an answer such as (17). This yielded a total of 623 *yes/no* questions.

(13) [sw00utt/sw_0018_4082.utt]

A: Do you, by mistakes, do you mean just like honest mistakes

A: or do you think they are deliberate sorts of things?

B: Uh, I think both.

(14) [sw02utt/sw_0220_2549.utt]

A: They finished pretty close to five hundred last year didn’t they?

B: Yeah.

(15) [sw02utt/sw_0496_3344.utt]

A: Senators are in for six years, right?

B: Right.

(16) [sw01utt/sw_0180_3134.utt]

A: That shouldn’t be too much trouble for the two of us. Right?

B: No.

(17) [sw00utt/sw_0070_3435.utt]

A: How do you feel about your game?

A: I guess that’s a good question?

B: Uh, well, I mean I’m not a serious golfer at all.

To identify indirect answers, I looked at the answer tags. The distribution of answers is given in Table 3.1. I collapsed the tags into 6 categories. Category I contains direct *yes/no* answers as well as "agree" answers (e.g., *That's exactly it*). Category II includes statement–opinion and statement–non-opinion: e.g., *I think it's great*, *Me I'm in the legal department*, respectively. Affirmative non-*yes* answers (e.g., *It is*) and negative non-*no* answers (e.g., *Uh, not a whole lot*) form category III. Other answers such as *I don't know* are in category IV. In category V, I put utterances that avoid answering the question: by holding (*I'm drawing a blank*), by returning the question — *wh*-question or rhetorical question (*Who would steal a newspaper?*) — or by using a backchannel in question form (*Is that right?*). Finally, category VI contains dispreferred answers, i.e., answers that disagree with, or decline, the positions of the interlocutor, as in (18) and (19) (Schegloff *et al.* 1977; Pomerantz 1984).

(18) [sw01utt/sw_0116_2406.utt]

A: Even better than Roger Rabbit insofar as animation?

B: Well, Roger was a composite one.

(19) [sw01utt/sw_0177_2759.utt]

A: That would probably bother me to wake up one day and find out halfway through the day that you're going to be drug tested and you didn't know about it.

B: Well, I guess it depends on if you got something to worry about.

A: I guess so. I guess so. Do you think that it's right?

B: Well, I come from kind of a biased opinion because I'm a, a therapist and a drug and alcohol counselor.

I hypothesized that indirect answers would appear in categories II, III and VI. However, some of the affirmative/negative non-*yes/no* answers ("na/ng") answers (category III in table 3.1) are disguised *yes/no* answers, such as *Right*, *I think so*, or *Not really*, and as such do not interest me. In the case of statements ("sv/sd") and

	Definition	Tag	Count
I	*yes/no* answers	ny/nn/aa	341
II	statements	sv/sd	143
III	affirmative/negative non-*yes/no* answers	na/ng	91
IV	other answers	no	21
V	avoid answering	^h/qw/qh/bh	18
VI	dispreferred answers	nd	9
Total			623

Table 3.1: Distribution of answer tags to *yes/no* questions in the Switchboard Dialog Act Corpus.

dispreferred answers ("nd"), many answers include reformulation, question avoidance (see 20), or a change of framing (21).

(20) [sw01utt/sw_0177_2759.utt]

A: Have you ever been drug tested?

B: Um, that's a good question.

(21) [sw00utt/sw_0046_4316.utt]

A: Is he the guy wants to, like, deregulate heroin, or something?

B: Well, what he wants to do is take all the money that, uh, he gets for drug enforcement and use it for, uh, drug education.

A: Uh-huh.

B: And basically, just, just attack the problem at the demand side.

I thus examined by hand all *yes/no* questions and found 88 examples of indirect answers (such as (22)–(25)), which constitutes 14% of the total answers to direct *yes/no* questions, a figure similar to the one in (Stenström 1984).

(22) [sw00utt/sw_0046_4316.utt]

A: That was also civil?

B: The other case was just traffic, and you know, it was seat belt law.

(23) [sw00utt/sw_0001_4325.utt]

A: Do you have kids?

B: I have three.

(24) [sw00utt/sw_0057_3506.utt]

A: Is it in Dallas?

B: Uh, it's in Lewisville.

(25) [sw01utt/sw_0160_3467.utt]

A: Are they [your kids] little?

B: I have a seven-year-old and a ten-year-old.

A: Yeah, they're pretty young.

3.2.2 Quantifying uncertain indirect answers

The next step was to assess how many of the answers to these 88 questions are non-committal (to use the term from Hockey *et al.* (1997)). In other words, how often inferring a 'yes' or 'no' answer to the question is unclear? I marked 32 as non-committal, i.e., 36.4% of the indirect answers I found, a figure similar to the 44% reported by Hockey *et al.* (1997).

For some of the question-answer pairs, determining the intended answer is straightforward. For example, in (23), the intended response is clearly 'yes'. B's reply is a case of "over-answer", i.e., a reply where more information is given than is strictly necessary to resolve the question. Hearers supply more information than strictly asked for when they recognize that the speaker's intentions are more general than the question posed might suggest. In (23), the most plausible intention behind the query is to know more about B's family. The hearer can also identify the speaker's plan and any necessary information for its completion, which he then provides (Allen & Perrault 1980). Other over-answers often require substantial amounts of linguistic and/or world knowledge to allow the inference to go through, as in (11) and (24). In the case of (11), one must recognize that *hysterical* is semantically stronger than *good*. Similarly, to recognize the implicit 'no' of (24), one must recognize that Lewisville is a distinct location from Dallas, rather than, say, contained in Dallas, and it must include more general constraints as well (e.g., an entity cannot be in two physical locations at once). However, once the necessary knowledge is in place, the inferences are properly licensed.

In other cases, the content of the answer itself does not fully resolve the question, known as "partially-resolved questions" (Groenendijk and Stokhof, 1984; Zeevat, 1994; Roberts, 1996; van Rooy, 2003). One instance is shown in (25), where the gradable adjective *little* is the source of difficulty. The response, while an answer, does not, in and of itself, resolve whether the children should be considered *little*. The predicate *little* is a gradable adjective, which inherently possesses a degree of vagueness: such adjectives contextually vary in truth conditions and admit borderline cases (Kennedy, 2007). In the case of *little*, while some children are clearly little, e.g., ages 2–3, and some clearly are not, e.g., ages 14–15, there is another class in between for which it is difficult to assess whether *little* can be truthfully ascribed to them. Due to the slippery nature of these predicates, there is no hard-and-fast way to resolve such questions in all cases. In (25), it is the questioner who resolves the question by accepting the information proffered in the response as sufficient to count as *little*. (26) is another example where the meaning of the adverb *often* is context dependent and how to resolve the question is therefore uncertain.

(26) [sw01utt/sw_0152_3108.utt]

A: Do you rent movies very often?

B: We generally rent a couple a week.

In (27), what counts as a crime will drive the inference. The definition of crime is however contextually dependent and can vary a lot, even when interpreted as a legal term. Note that the disjunct *or anything* may indicate that A is open to hearing about alternatives to *stealing* as definition of *crime*.

(27) [sw00utt/sw_0048_4340.utt]

A: Have you been the subject of such a crime, such as stealing, or anything?

B: I've been caught with marijuana before.

B: So I guess that was a crime.

Polysemy provides another difficult case. In (23), we saw an indirect answer to the question *Do you have kids?* which was a clear case of entailment; (28) shows a more complicated indirect answer, which is at once positive, the speaker asserts that she has children, viz. biological descendants, and negative, the speaker asserts that her children are no longer children in terms of age. Had B responded with a simple *yes*, A would have wrongly concluded that B had children in both senses.

(28) [sw00utt/sw_0189_3266.utt]

A: Do you have any children?

B: Uh, they're all grown up.

The corpus study confirms the pervasiveness of indirect answers to polar questions and emphasizes that such answers may not always be definitively resolved to 'yes' or 'no'.

3.3 Previous work

Indirect answers to polar questions have already been investigated in the computational literature. For instance, Green and Carberry (1994, 1999) provide an extensive model that interprets and generates such answers. Their model is based on the rhetorical structure theory (RST) of Mann & Thompson (1988). RST provides a set of relations to describe the organization of coherent text. Relations (e.g., condition,

Answer-yes	**Answer-no**
Use-condition	Use-otherwise
Use-elaboration	Use-obstacle
Use-cause	Use-contrast
Use-elaboration	**Use-obstacle**
Use-cause	Use-obstacle
Use-elaboration	Use-elaboration

Figure 3.1: Discourse plan operators for *yes* and *no* answers, as well as for two satellite discourse plan operators.

contrast, elaboration, purpose) are defined between two pieces of text, called the nucleus and the satellite. In (29) from Green & Carberry (1999), a contrast relation holds between the question (treated as the nucleus) and the answer (the satellite).

(29) A: Aren't you going shopping tonight?

B: I'm going tomorrow night.

Green and Carberry's model links such coherence relations to discourse plan operators which encode generic programs of communicative actions. Discourse plan operators are thus defined for 'yes' and 'no' answers, schematically represented in Figure 3.1. A 'yes' answer can be implemented by three different types of answer (or satellites), which are themselves discourse plan operators: use-condition, use-elaboration or use-cause; a 'no' answer can be implemented by three other satellites: use-otherwise, use-obstacle or use-contrast. The satellites can all be present or some only, in any order. Figure 3.1 also displays the use-elaboration and use-obstacle discourse plan operators. Such discourse plan operators can be used to interpret and generate answers. In (30), the identification of an obstacle relation between the question and the answer leads to the interpretation of the answer as 'no' since the obstacle relation appears under the answer-no operator. On the other hand in (31), an elaboration relation leads to the interpretation of the answer as 'yes'.

(30) A: Are you going to campus tonight?

B: My car is broken.

(31) A: Do you collect classic automobiles?
B: I recently purchased an Austin-Healey 3000.

A strong point of Green and Carberry's model is its ability to deal with multi-utterance responses. The model takes into account subsequent discourse context, and can therefore provide the correct analysis for (32) and (33).

(32) A: Are you going shopping?
B: I'm going to campus.
B: I have a night class.

(33) A: Are you going shopping?
B: I'm going to campus.
B: The bookstore is having a sale.

In (33), the presence of B's second utterance *The bookstore is having a sale* will lead the system to analyze B's response as a use-elaboration satellite instantiated by both an elaboration relation *I'm going to campus* and a cause relation *The bookstore is having a sale*. The response will thus be interpreted as positive: use-elaboration appears under the answer-yes operator. On the other hand, B's response in (32) will be analyzed as a use-obstacle satellite instantiated by an obstacle relation as well as an elaboration relation, and therefore interpreted as a 'no' answer.

Thus Green and Carberry propose a logical inference model which makes use of discourse plan operators and coherence relations to infer *categorical* answers. However as emphasized in the previous section, to adequately interpret indirect answers, the uncertainty inherent in some answers needs to be captured. Another disadvantage of their method is the amount of hard-coded knowledge that is required: all the discourse plan operators need to be detailed as well as knowledge about the world talked about, e.g., knowing that you shop in bookstores but not in night classes. Their system is evaluated on a very specific setting: questions are asked about a few different objects placed on and around a table (a blue block on top of a red block which is placed on the table, a yellow ball next to the table on the ground, etc.). Such a limited setting enables them to encode the necessary world knowledge.

Here, I propose to handle the uncertainty inherent in the interpretation of some indirect answers to *yes/no* questions. I also aim at a system with broad coverage. I learn knowledge that can be extracted from large amounts of text instead of coding for specific domain knowledge.

3.4 Gradable adjectives

Section 3.2 highlighted the importance of capturing the uncertainty that can be conveyed in indirect answers. This uncertainty is especially pressing for predications built around scalar modifiers which can be vague and highly context-dependent (Kamp & Partee 1995; Kennedy & McNally 2005; Kennedy 2007). Such modifiers, also called gradable adjectives, will be the target of my study.

Gradable adjectives, e.g., *tall* or *wet*, are modifiers that participate in a scale: for example, *minuscule* < *short* < *tall* < *gigantic* or *damp* < *wet* < *drenched*. They come in two categories: relative and absolute gradable adjectives (Kennedy 2007). Kennedy (2007) argues that relative gradable adjectives are inherently vague whereas absolute gradable adjectives are not. The meaning of relative gradable adjectives, like *tall* or *expensive*, depends on a standard of comparison which varies according to the context: someone can be tall for a gymnast, but not for a basketball player. Such relative gradable adjectives also allow for borderline cases. For example, even if we fix the basic sense for *little* to mean ‘young for a human’, there is a substantial amount of gray area between the clear instances (babies) and the clear non-instances (adults). This is the source of uncertainty in (25) above, in which B’s children (7 and 10 year-old) fall into the gray area. On the other hand, absolute gradable adjectives, such as *empty*, participate in a scale but do not have context dependent interpretations and do not exhibit borderline cases. There are two varieties of absolute gradable adjectives: minimum standard and maximum standard. Minimum standard absolute adjectives require their arguments to possess some minimal degree of the property they describe. Maximum standard absolute adjectives require their arguments to possess a maximal degree of the property: *the glass is empty* means that it does not contain anything. The absolute gradable adjectives are thus not vague per se as relative gradable adjectives are, though they allow for some imprecision. *Empty* can be used to describe a situation where there is a little bit of content in the container, instead of nothing at all: *The theater is empty tonight* is not infelicitous when used to talk about a theatre where only a very few people are present.

Gradable adjectives provide a perfect case study for my purposes. Since such adjectives participate in a scale, they can easily give rise to indirect answers. For example, if asked whether I am happy, instead of uttering a direct *yes* or *no* to answer the question, I can choose to use another adjective on the relevant scale (e.g., *cheerful*, *thrilled*, *overjoyed*), therefore offering an indirect answer to the question. Moreover, the interpretation of some answers is inherently uncertain due to the nature of gradable adjectives. The intended 'yes' or 'no' response is clear in some indirect answers involving gradable adjectives, as in (34), but the answer is less certain in other cases, as in (35):

(34) [LouDobbsTonight/0410-19-ldt.01.txt]

A: Do you think that's a good idea, that we just begin to ignore these numbers?

B: I think it's an excellent idea.

(35) [TheSituationRoom/0703-16-sitroom.01.txt]

A: Is Obama qualified?

B: I think he's young.

3.5 Corpus description

In this chapter, I focus on interpreting indirect answers to questions involving gradable adjectives. To evaluate the learning techniques for interpreting such question-answer pairs, I gathered instances of those pairs. Since indirect answers are likely to arise in interviews, I used online CNN interview transcripts from 5 different shows aired between 2000 and 2008 (Anderson Cooper, Larry King Live, Late Edition, Lou Dobbs Tonight, The Situation Room). I used regular expressions and manual filtering to find examples of two-utterance dialogues in which the question and the reply contain some kind of gradable modifier. I also used a subset of the data from the Switchboard Dialog Act corpus that I analyzed in section 3.2.1.

3.5.1 Types of question-answer pairs

In total, I ended up with 224 question-answer pairs involving gradable adjectives. However the collection contains different types of answers, which naturally fall into two categories: (I) in 205 dialogues, both the question and the answer contain a gradable modifier; (II) in 19 dialogues, the reply contains a numerical measure (as in (12), (25) and (36)).

(36) [sw09utt/sw_0964_2237.utt]

A: Do you have a long growing season up there?

B: We have about three months.

Category (I), which consists of pairs of modifiers, can be further divided. In most dialogues, the answer contains an adjective other than the one used in the question, such as in (34). In others, the answer contains the same adjective as in the question, but modified by an adverb (e.g., *very*, *basically*, *quite*) as in (37) or a negation as in (38).

(37) [AndersonCooper360/0705-30-acd.02.txt]

A: That seems to be the biggest sign of progress there. Is that accurate?

B: That's absolutely accurate.

(38) [LarryKingLive/0601-20-lkl.01.txt]

A: Are you bitter?

	Modification in answer	Example	Count
I	Other adjective	(34), (35)	125
	Adverb - same adjective	(37)	55
	Negation - same adjective	(38), (39)	21
	Omitted adjective	(40)	4
II	Numerical measure	(25), (36)	19

Table 3.2: Types of question-answer pairs, and counts in the corpus.

B: I'm not bitter because I'm a soldier.

The negation can be present in the main clause when the adjectival predication is embedded, as in (39).

(39) [AndersonCooper360/0705-14-acd.01.txt]

A: [. . .] Is that fair?

B: I don't think that's a fair statement.

In a few cases, when the question contains an adjective modifying a noun, the adjective is omitted in the answer:

(40) [LateEdition/0207-07-le.00.txt]

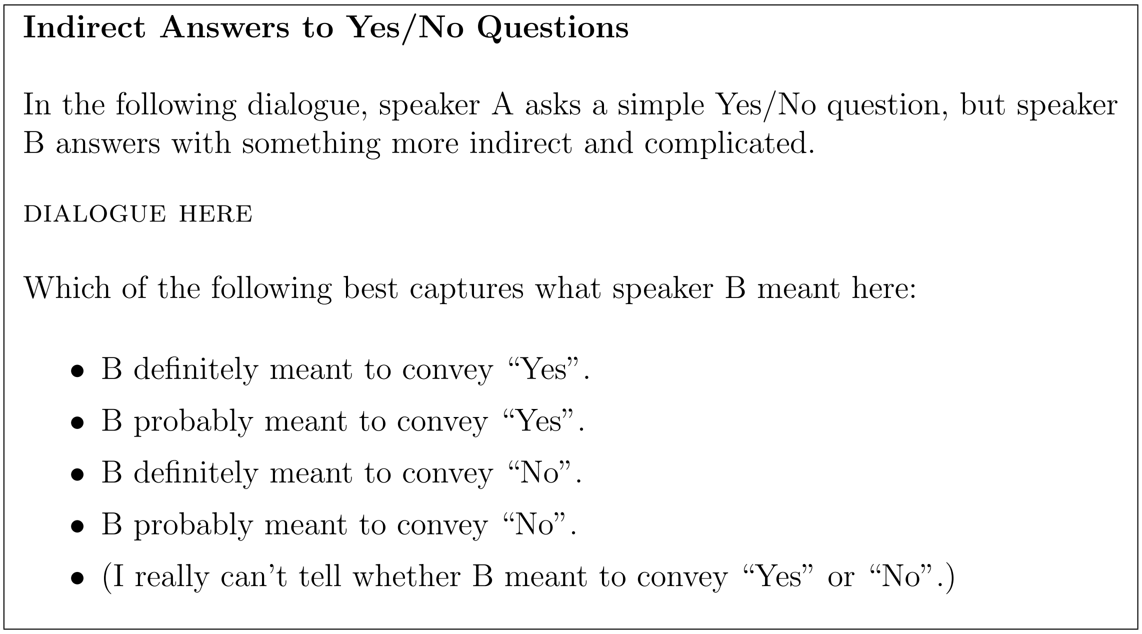
A: Is that a huge gap in the system?

B: It is a gap.

Table 3.2 gives the distribution of the types appearing in the corpus.

3.5.2 Answer assignment

To assess the degree to which each answer conveys 'yes' or 'no' in context, I used Amazon's Mechanical Turk (AMT) service, and gathered response distributions from the Mechanical Turk subjects (Turkers, in AMT parlance). Given a written dialogue between speakers A and B, Turkers were asked to judge what B's answer conveys: 'definite yes', 'probable yes', 'uncertain', 'probable no', 'definite no'. Figure 3.2 shows the exact formulation used in the experiment.



Indirect Answers to Yes/No Questions

In the following dialogue, speaker A asks a simple Yes/No question, but speaker B answers with something more indirect and complicated.

DIALOGUE HERE

Which of the following best captures what speaker B meant here:

- B definitely meant to convey “Yes”.
- B probably meant to convey “Yes”.
- B definitely meant to convey “No”.
- B probably meant to convey “No”.
- (I really can’t tell whether B meant to convey “Yes” or “No”.)

Figure 3.2: Design of the Mechanical Turk experiment.

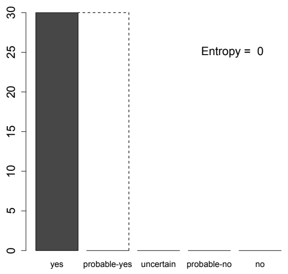
For each dialogue, I got answers from 30 Turkers, and took the dominant response as the correct one.[2] I also computed entropy values for the distribution of answers for each item. Overall, the agreement was good: 21 items have total agreement (entropy of 0.0 — 11 in the “adjective” category, 9 in the “adverb-adjective” category and 1 in the “negation” category), and 80 items are such that a single response got chosen 20 or more times (entropy < 0.9). The dialogues in (34) and (41) are examples of total agreement. In contrast, (42) has response entropy of 1.1, and item (43) has the highest entropy of 2.2. Next to the dialogues, I display the exact Turkers’ distribution. The dotted lines pool together the positive (‘yes’ and ‘probable yes’) and negative (‘probable no’ and ‘no’) answers.

[2] 120 Turkers were involved (the median number of items done was 28 and the mean 56.5). The Fleiss’ Kappa score for the five response categories is 0.46, though these categories are partially ordered. For the three-category response system used in section 3.7, which arguably has no scalar ordering, the Fleiss’ Kappa is 0.63. When using the three-category response system, 66 items have a 0.0 entropy, meaning that the Turkers either do not disagree at all on the judgments or only disagree in terms of whether the answer is ‘probable’ or ‘definite’. Despite variant individual judgments, aggregate annotations done with Mechanical Turk have been shown to be reliable (Snow *et al.* 2008; Sheng *et al.* 2008). Here, the relatively low Kappa scores also reflect the uncertainty inherent in many of the examples, uncertainty that I seek to characterize and come to grips with computationally.

(41) [LarryKingLive/0808-28-lkl.01.txt]

A: You said it was an advertisement. Advertisements can be good or bad. Was it a good ad?

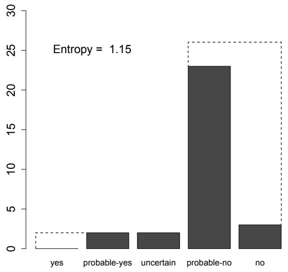
B: It was a great ad.



(42) [LouDobbsTonight/0410-19-ldt.01.txt]

A: Am I clear?

B: I wish you were a little more forthright.



(43) [LateEdition/0109-22-le.00.txt]

A: 91 percent of the American people still express confidence in the long-term prospect of the U.S. economy; only 8 percent are not confident. Are they overly optimistic, in your professional assessment?

B: I think it shows how wise the American people are.

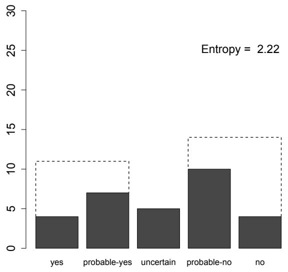


Table 3.3 shows the mean entropy values for the different categories identified in the corpus.

Interestingly, the pairs involving an adverbial modification in the answer all received a positive answer (‘yes’ or ‘probable yes’) as dominant response. All 19 dialogues involving a numerical measure had either ‘probable yes’ or ‘uncertain’ as dominant response. There is thus a significant bias for positive answers: 70% of the

I	Modification in answer	Mean	SD
	Other adjective	1.1	0.6
	Adverb - same adjective	0.8	0.6
	Negation - same adjective	1.0	0.5
	Omitted adjective	1.1	0.2
II	Numerical measure	1.5	0.2

Table 3.3: Mean entropy values and standard deviation obtained in the Mechanical Turk experiment for each question-answer pair category.

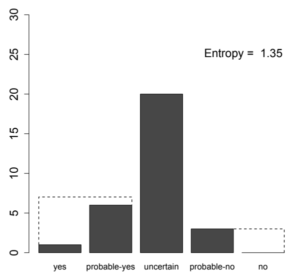
category I items and 74% of the category II items have a positive answer as dominant response. Examining the subset of the Dialog Act corpus described in section 3.2.1, I found that 38% of the *yes/no* questions receive a direct positive answers, whereas 21% have a direct negative answer. This bias probably stems from the fact that people are more likely to use an overt denial expression where they need to disagree, whether or not they are responding indirectly.

Given the set-up of the Mechanical Turk experiment, uncertainty in the judgments about the relationship between the adjectives can arise in two ways: the dominant response from the Turkers falls in the 'uncertain' category or there is a high entropy in the Turkers' responses. Below I show examples of such cases, with the response distribution obtained from the Turkers. The examples reported here give a flavor of the kind of uncertainty that is carried in conversation.[3]

(44) [LarryKingLive/0005-10-lkl.00.txt]

A: This is not just one party or the other. Is the whole thing almost like hypocritical?

B: Well, I think it's interesting.

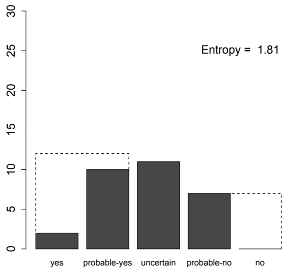


[3] In some examples, the modifier is itself modified (e.g., *almost like hypocritical*, *roughly half an acre*. Even though such modification might have a role to play, I have not addressed it here.

(45) [sw01utt/sw_0160_3467.utt]

A: Was she the best one on that old show?

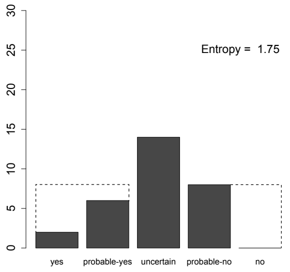
B: She was just funny.



(46) [sw00utt/sw_0087_2775.utt]

A: Do you have a big lot or anything like that?

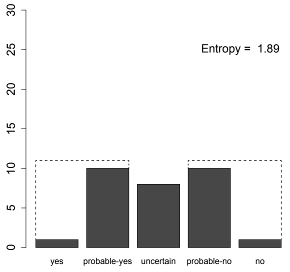
B: The whole lot I'm sitting on is roughly half an acre.



(47) [AndersonCooper360/0312-12-acd.00.txt]

A: Well, it's just one clip. Was it a good movie?

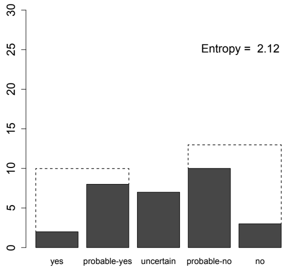
B: It's an OK movie.



(48) [LarryKingLive/0301-06-lkl.00.txt]

A: Just asked about alcohol. Does it have a large impact?

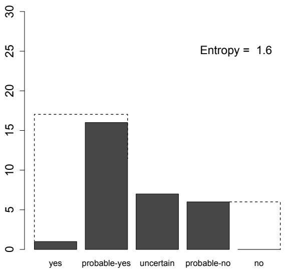
B: It has a medium-sized impact.



(49) [sw05utt/sw_0598_2858.utt]

A: Are you newly married?

B: Two years.

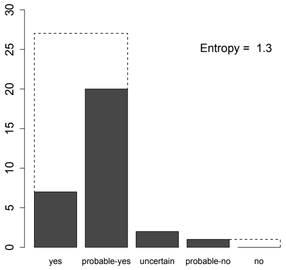


The uncertain judgment in (46) probably stems from the absence of context. In that excerpt there is no mention of the place talked about, and it is therefore difficult to come up with a standard of comparison: out of the blue, it is unclear whether half an acre is considered to be a big lot or not. Other numerical examples, on the other hand, intrinsically carry a standard of comparison, and since the numerical measures given in the answers are not borderline cases, the Turkers agree with each other: in (50), there is some consensus that eighty degrees is considered warm weather; in (51), five and two year-olds are considered to be young kids.

(50) [Fisher/fe_03_05856.txt]

A: Is the weather warm there?

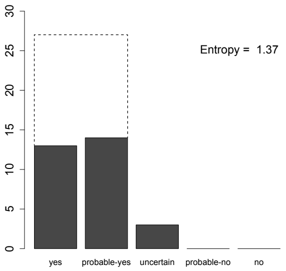
B: We've had almost eighty degrees for this week.



(51) [sw04utt/sw_0489_3238.utt]

A: Are your kids young?

B: Five and two year-old.



3.6 Methods

I now turn to the methods I propose for grounding the meanings of scalar modifiers. I develop two methods which target the two categories of answers I found: category I where both the question and the answer contain a gradable modifier, and category II where the answer contains a numerical measure. The AMT data described in the previous section will serve to evaluate the techniques proposed here.

3.6.1 Learning modifier scales and inferring *yes/no* answers

The first technique targets items where both the question and the answer contain a gradable modifier (category I in the corpus). The central hypothesis is that, in polar question dialogues, the semantic relationship between the main predication P_Q in the question and the main predication P_A in the answer is the primary factor in determining whether, and to what extent, 'yes' or 'no' was intended. If P_A is at least as strong as P_Q, the intended answer is 'yes'; if P_A is weaker than P_Q, the intended answer is 'no'; and, where the relative ordering between P_A and P_Q is unclear, the answer is uncertain.

For example, *good* is weaker than *excellent* (i.e., on the relevant scale between *good* and *excellent*, *excellent* will appear after *good*), and thus speakers infer that the reply in example (34) above is meant to convey 'yes'. In contrast, if we reverse the order of the modifiers — roughly, *Is it a great idea?*; *It's a good idea* — then speakers infer that the answer conveys 'no'. Had B replied with *It's a complicated idea* in (34), then uncertainty would likely have resulted, since *good* and *complicated* are not in a reliable scalar relationship. It is also important to take negation into account. Negation reverses scales (Horn 1972; Levinson 2000), so it flips 'yes' and 'no' in these cases, leaving 'uncertain' unchanged. When both the question and the answer contain a modifier (such as in (41)–(43)), the *yes/no* response should correlate with the extent to which the pair of modifiers can be put into a relative order along a relevant scale.

To learn such scales from text, I collected a large corpus of online reviews from the Internet Movie Database (IMDB). Each of the reviews in this collection has an

Rating	Reviews	Words	Vocabulary	Average words per review
1	124,587	25,389,211	192,348	203.79
2	51,390	11,750,820	133,283	228.66
3	58,051	13,990,519	148,530	241.00
4	59,781	14,958,477	156,564	250.22
5	80,487	20,382,805	188,461	253.24
6	106,145	27,408,662	225,165	258.22
7	157,005	40,176,069	282,530	255.89
8	195,378	48,706,843	313,046	249.30
9	170,531	40,264,174	273,266	236.11
10	358,441	73,929,298	381,508	206.25
Total	1,361,796	316,956,878	1,160,072	206.25

Table 3.4: Numbers of reviews, words and vocabulary size per rating category in the IMDB review corpus, as well as the average number of words per review.

associated star rating: one star (most negative) to ten stars (most positive). Table 3.4 summarizes the distribution of reviews as well as the number of words and vocabulary across the ten rating categories.

As is evident from table 3.4, there is a significant bias for ten-star reviews. This is a common feature of such corpora of informal, user-provided reviews (Chevalier & Mayzlin 2006; Hu *et al.* 2006; Pang & Lee 2008). However, since I do not want to incorporate the linguistically uninteresting fact that people tend to write a lot of ten-star reviews, I assume uniform priors for the rating categories. Let $\text{count}(w, r)$ be the number of tokens of word w in reviews in rating category r, and let $\text{count}(r)$ be the total word count for all words in rating category r. The probability of w given a rating category r is simply $\Pr(w|r) = \text{count}(w, r)/\,\text{count}(r)$. Then under the assumption of uniform priors, we get $\Pr(r|w) = \Pr(w|r)/\sum_{r' \in R} \Pr(w|r')$.

In reasoning about the dialogues, I rescale the rating categories by subtracting 5.5 from each, to center them at 0. This yields the scale $R = \langle -4.5, -3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5, 4.5\rangle$. The rationale for this is that modifiers at the negative end of the scale (*bad*, *awful*, *terrible*) are not linguistically comparable to those at the positive end of the scale (*good*, *excellent*, *superb*). Each group forms its own qualitatively different scale (Kennedy & McNally 2005). Rescaling allows us to make a basic positive vs. negative distinction. Once done, an increase in absolute value is an increase in strength. In the experiments, I use expected rating values to characterize the polarity and strength of modifiers. The expected rating value for a word w is $\text{EV}(w) = \sum_{r \in R} r \Pr(r|w)$. This single statistic $\text{EV}(w)$ is sufficient to place adjectives into scales. Figure 3.3 plots these values for a number of scalar terms, both positive and negative, across the rescaled ratings, with the vertical lines marking their EV values. The weak scalar modifiers all the way on the left are most common near the middle of the scale, with a slight positive bias in the top row and a slight negative bias in the bottom row. As we move from left to right, the bias for one end of the scale grows more extreme, until the words in question are almost never used outside of the most extreme rating category. The resulting scales correspond well with linguistic intuitions and thus provide an initial indication that the rating categories are a reliable guide to strength and polarity for scalar modifiers. I put this information to use in the corpus via the decision procedure described in figure 3.4.

3.6.2 Interpreting numerical answers

The second technique aims at determining whether a numerical answer counts as a positive or negative instance of the adjective in the question (category II in the corpus).

Dimensional adjectives that can receive attributes describes in terms of measure, such as *little* or *long*, inherently possess a degree of vagueness (Kamp & Partee 1995; Kennedy 2007): while in the extreme cases, judgments are strong (e.g., *a six foot tall woman* can clearly be called “a tall woman” whereas *a five foot tall woman* cannot), there are borderline cases for which it is difficult to say whether the adjectival predication can truthfully be ascribed to them. A logistic regression model can capture these observations. To build this model, I gather distributional information from the Web.

For instance, in the case of (25), I can retrieve from the Web positive and negative examples of age in relation to the adjective and the modified entity *little kids*. The

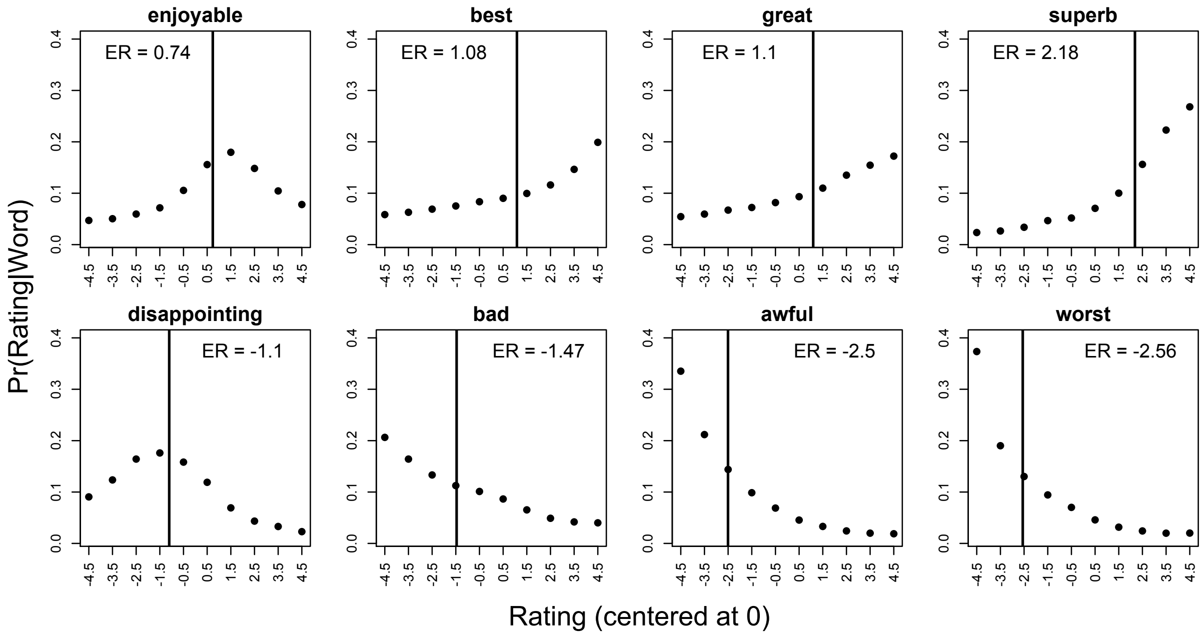


Figure 3.3: The distribution of some scalar modifiers across the ten rating categories. The vertical lines mark the expected ratings, defined as a weighted sum of the probability values (black dots).

Let D be a dialogue consisting of
(i) a polar question whose main predication is based on scalar predicate P_Q and
(ii) an indirect answer whose main predication is based on scalar predicate P_A.
Then:

1. if P_A or P_Q is missing from our data, infer ‘Uncertain’;
2. else if EV(P_Q) and EV(P_A) have different signs, infer ‘No’;
3. else if abs(EV(P_Q)) $\leqslant$ abs(EV(P_A)), infer ‘Yes’;
4. else infer ‘No’.
5. In the presence of negation, map ‘Yes’ to ‘No’, ‘No’ to ‘Yes’, and ‘Uncertain’ to ‘Uncertain’.

Figure 3.4: Decision procedure for using the word frequencies across rating categories in the review corpus to decide what a given answer conveys.

question contains the adjective and the modified entity. The reply contains the unit of measure (here *year-old*) and the numerical answer. Specifically I query the Web using Yahoo! BOSS (Academic) for *"little kids" year-old* (positive instances) as well as for *"not little kids" year-old* (negative instances). Yahoo! BOSS is an open search services platform that provides a query API for Yahoo! Web search. I then extract ages from the positive and negative snippets obtained, and fit a logistic regression to these data. To remove noise, I discard low counts (positive and negative instances for a given unit < 5). Also, for some adjectives, such as *little* or *young*, there is an inherent ambiguity between absolute and relative uses. Ideally, a word sense disambiguation system would be used to filter these cases. For now, I extract the largest contiguous range for which the data counts are over the noise threshold. Otherwise, the model is ruined by references to "young 80-year olds", using the relative sense of *young*, which are moderately frequent on the Web. When not enough data is retrieved for the negative examples, I expand the query by moving the negation outside the search phrase. I also replace the negation and the adjective by the antonyms given in WordNet (using the first sense).

The logistic regression thus has only one factor — the unit of measure (age in the case of *little kids*). For a given answer, the model assigns a probability indicating the extent to which the adjectival property applies to that answer. If the factor is a significant predictor, I can use the probabilities from the model to decide whether the answer qualifies as a positive or negative instance of the adjective in the question, and thus interpret the indirect response as a 'yes' or a 'no'. The probabilistic nature of this technique dovetails with the inherent uncertainty carried by some indirect answers.

3.7 Evaluation and results

My primary goal is to evaluate how well I can learn the relevant scalar and entailment relationships from the Web. In the evaluation, I thus applied the techniques proposed to a manually coded version of the 224 question-answers pairs corpus. For the adjectival scales, I annotated each example for its main predication (modifier, or adverb–modifier bigram), including whether that predication was negated. I also manually coded for affixal negation of the adjectives (e.g., *unconfirmed*).[4] For the numerical cases, I manually constructed the initial queries. However, identifying the requisite predications and recognizing the presence of negation or embedding could be done automatically using dependency graphs. As a test, I transformed the corpus into the Stanford dependency representation (see chapter 2), and was able to automatically retrieve all negated modifier predications, except for one (*We had a view of it, not a particularly good one*), where a parse error leads to wrong dependencies.

To evaluate the techniques, I pool the Mechanical Turk 'definite yes' and 'probable yes' categories into a single category 'Yes', and do the same for 'definite no' and 'probable no'. Together with 'uncertain', this makes for three-response categories. I count an inference as successful if it matches the dominant Turker response category. To use the three-response scheme in the numerical experiment, I simply categorize the probabilities as follows: 0–0.33 = 'No', 0.33–0.66 = 'Uncertain', 0.66–1.00 = 'Yes'.

Table 3.5 gives a breakdown of the system's performance on the various category subtypes. The overall accuracy level is 71% (159 out of 224 inferences correct). Table 3.6 summarizes the results per response category, for the examples in which both the question and answer contain a gradable modifier (category I), and for the numerical cases (category II).

3.8 Error analysis

Accuracy is high on the "Adverb – same adjective" and "Negation – same adjective" cases because the 'Yes' answer is fairly direct for them (though adverbs like *basically* introduce an interesting level of uncertainty). The system is less accurate for the "Other adjective" category.

Inferring the relation between scalar adjectives has some connection with work in

[4] To deal with such adjectives, I opted to compare the non-negated forms, and then flip the answer as outlined above in figure 3.4. For example, if the dialogue involves the adjectives *sad* and *unhappy*, the system compares the ER values for *sad* and *happy*, and then reverses the answer obtained. Another approach would be to directly compare the adjectives as is. There were only five dialogues involving affixal negations. Post-hoc analyses show that for these items, both methods would have lead to the same results.

	Modification in answer	Precision	Recall
I	Other adjective	60	60
	Adverb - same adjective	95	95
	Negation - same adjective	100	100
	Omitted adjective	100	100
II	Numerical	89	40
Total		75	71

Table 3.5: Summary of precision and recall (%) by type.

	Response	Precision	Recall	F1
I	Yes	87	76	81
	No	57	71	63
II	Yes	100	36	53
	Uncertain	67	40	50

Table 3.6: Precision, recall, and F1 (%) per response category. In the case of the scalar modifiers experiment, there were just two examples whose dominant response from the Turkers was ‘Uncertain’, so this category is left out of the results. In the numerical experiment, there were no ‘No’ answers.

	Response	Precision	Recall	F1
WordNet-based	Yes	82	83	82.5
(items I)	No	60	56	58

Table 3.7: Precision, recall, and F1 (%) per response category for the WordNet-based approach.

sentiment detection. Even though most of the research in that domain focuses on the orientation of one term using seed sets, techniques which provide the orientation strength could be used to infer a scalar relation between adjectives. For instance, Blair-Goldensohn *et al.* (2008) use WordNet to develop sentiment lexicons in which each word has a positive or negative value associated with it, representing its strength. The algorithm begins with seed sets of positive, negative, and neutral terms, and then uses the synonym and antonym structure of WordNet to expand those initial sets and refine the relative strength values. Using my own seed sets, I built a lexicon using Blair-Goldensohn *et al.*'s (2008) method and applied it as in figure 3.4 (changing the EV values to sentiment scores). Both approaches achieve similar results: for the "Other adjective" category, the WordNet-based approach yields 56% accuracy, which is not significantly different from the performance of the technique I suggested (60%); for the other types in category I, there is no difference in results between the two methods. Table 3.7 summarizes the results per response category for the WordNet-based approach (which can thus be compared to the category I results in table 3.6). However in contrast to the WordNet-based approach, my method requires no hand-built resources: the synonym and antonym structures, as well as the strength values, are learned from Web data alone. In addition, the WordNet-based approach must be supplemented with a separate method for the numerical cases.

In the "Other adjective" category, 31 items involve opposites (Cruse 1986; Cruse 2011; Lehrer & Lehrer 1982). Opposites are terms that lie in an inherently incompatible binary relationship. Cruse (1986) identifies three basic properties of opposites: (i) there are only two members of a set of opposites, so the relationship must be binary; (ii) the binarity has to be inherent (*coffee or tea?* is a standard question, but the binarity there is pragmatic and not inherent, in contrast with terms such as *up* and *down*); (iii) the binarity has to be patent, i.e., be a salient part of the meaning of the words (for example, *yesterday* and *tomorrow* are opposites but not *Monday* and *Wednesday*: the opposite directionality of *yesterday* and *tomorrow* relative to *today* is encoded in their meanings, whereas the opposite directionality of *Monday* and *Wednesday* relative to *Tuesday* has to be inferred). Different classifications of opposites have been proposed, one of which can be found in Cruse (1986) and Cruse

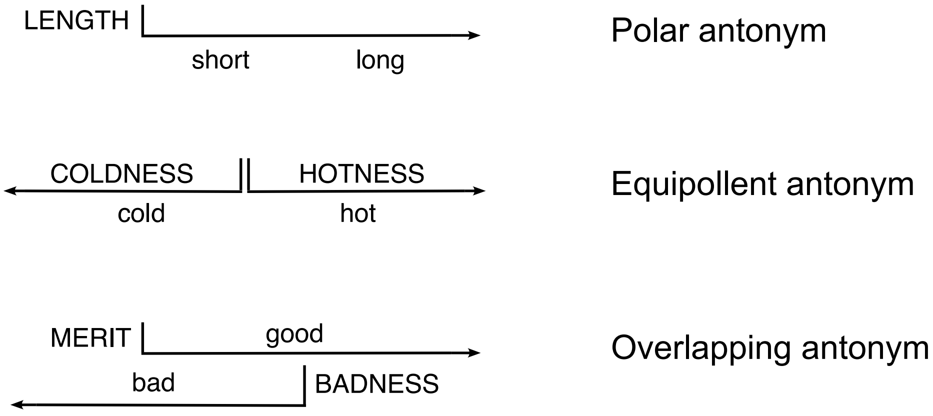


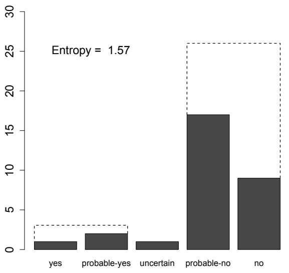
Figure 3.5: Representation of the scales for the different antonym classes (from Cruse & Togia 1995).

(2011). Cruse (2011) divides opposites into four principal classes: complementaries (*true*/*false*, *dead*/*alive*) where the two terms are in a mutually exclusive relationship; reversives (*up*/*down*, *dress*/*undress*) which denote movement in opposite directions or a change in opposite directions between states; converses (*husband*/*wife*, *doctor*/*patient*, *above*/*below*) which, contrary to all other classes of opposites, can describe the same situation (A is B's wife and B is A's husband); and antonyms (*long*/*short*, *hot*/*cold*). Antonyms are the class that interests us since they are gradable. The class of antonyms is subdivided into three groups: polar antonyms (*long*/*short*), equipollent antonyms (*hot*/*cold*) and overlapping antonyms (*good*/*bad*). Antonyms can be thought of as operating over scales representing values of some gradable property, such as length (*long* and *short*), temperature (*hot* and *cold*), or merit (*good* and *bad*). The different types of antonyms are related to different arrangements of the scales as represented in Figure 3.5. The 31 items in the dataset are mainly antonyms. Most items are standard antonyms that are found in antonym lists (e.g., *right*/*wrong*, *good*/*bad*), but some pairs, such as *ready*/*premature*, *true*/*preposterous*, *qualified*/*young*, are contingent antonyms as they depend on the context. There are also a few complementaries (*possible*/*impossible*).

The technique presented here accurately deals with a broad range of opposites, even though the technique does not explicitly model such terms. Opposites are at the ends of the same scale, and the technique assigns them negative and positive EV values. The technique finds prototypical opposites, but also opposites such as *confident/ nervous* (52), *acceptable/unprecedented* (53) that are lacking in antonymy resources (e.g., WordNet) or automatically generated antonymy lists (Mohammad *et al.* 2008; Mohammad *et al.* 2011). Mohammad *et al.*'s method uses an affix-generated seed set to compile a list of opposites from a thesaurus. Opposites created with affixes (*clear/unclear*, *honest/dishonest*) are looked for in the thesaurus categories: when such opposites are found across two categories, all the pairs of words in the two categories are considered to be opposites. The only manual part in this method is the creation of 15 patterns to create opposites via affixation, but the approach relies on the existence of a thesaurus. Mohammad *et al.* propose a second method which also makes use of the structure of the thesaurus: because of conventions of the thesaurus, adjacent categories in it tend to be contrasting ones. Adjacent categories were checked manually to make sure that they were in contrast with each other. All pairs between adjacent categories are considered opposites.

(52) [AndersonCooper360/0601-31-acd.01.txt]

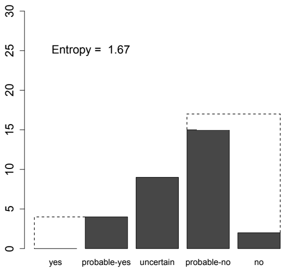
A: Since 9/11, whenever the president has argued with the Democrats on issue of terrorism, he almost always has prevailed. Are Republicans confident that, in this election year, he can help them prevail, as well?



B: Republicans are nervous.

(53) [LateEdition/0505-15-le.01.txt]

A: If the Democrats, Senator Lugar, were to go ahead and use that filibuster rule to try to set back the Bolton nomination, would that be unprecedented? Would that be acceptable?



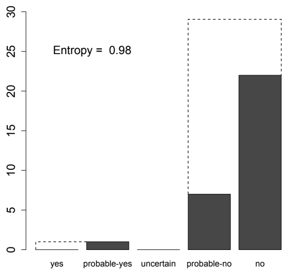
B: Well, it wouldn't be unprecedented.

Out of the 31 items in the dataset involving opposites, the technique presented here correctly marks 18, whereas Blair-Goldensohn *et al.*'s (2008) technique finds 11, and Mohammad *et al.*'s (2011) list compiled with the affix-generated seed set would correctly mark 13. The list of contrasting pairs that results from combining both methods in Mohammad *et al.* (2011) is very comprehensive and would correctly mark 20 items. However the technique presented here has the advantage of not requiring any hand-built resources. Right now the technique is solely based on unigrams, but could be improved by adding context: making use of dependency information, as well as moving beyond unigrams. The two dialogues in (54) and (55) involve the same opposites (*lot*/*few*), but received different responses from the Turkers. The current method fails to make a distinction between the two, since only the opposites are taken into account. However the adverb modifying *few* is a crucial component in the interpretation.

(54) [sw03utt/sw_0373_4013.utt]

A: Do you have a lot of large families out there in Texas?

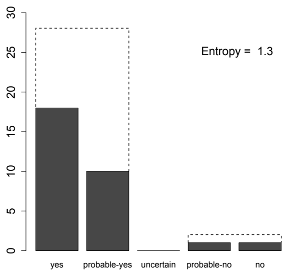
B: Very few.



(55) [sw04utt/sw_0404_2667.utt]

A: Well, do a lot of people take advantage of it?

B: Quite a few do.



In the numerical cases, precision is high but recall is low. For roughly half of the items, not enough negative instances can be gathered from the Web and the model lacks predictive power (as for items (36) or (56)).

(56) [sw00utt/sw_0007_4171.utt]

A: Do you happen to be working for a large firm?

B: It's about three hundred and fifty people.

Looking at the negative hits for item (56), one sees that few give an indication about the number of people in the firm, but rather qualifications about colleagues or employees (*great people*, *people's productivity*), or the hits are less relevant: "Most of the *people* I talked to were actually pretty optimistic. They were rosy on the job market and many had jobs, although most were *not large firm* jobs". The lack of data comes from the fact that the queries are very specific, since the adjective depends on the item modified (e.g., "expensive exercise bike", "deep pond"). However when we do get a predictive model, the probabilities correlate almost perfectly with the Turkers' responses. This happens for 8 items: "expensive to call (50 cents a minute)", "little kids (7 and 10 year-old)", "long growing season (3 months)", "lot of land (80 acres)", "warm weather (80 degrees)", "young kids (5 and 2 year-old)", "young person (31 year-old)" and "large house (2,450 square feet)". In the latter case only, the system output doesn't correlate with the Turkers' judgment. The system replies 'uncertain' where the dominant answer from the Turkers is 'probable yes' with 15 responses (but 11 answers from the Turkers are 'uncertain').

The logistic curves in figure 3.6 capture nicely the intuitions that people have about the relation between age and "little kids" or "young kids", as well as between Fahrenheit degrees and "warm weather". For "little kids", the probabilities of being little or not are clear-cut for ages below 7 and above 15, but there is a region of vagueness in between. In the case of "young kids", the probabilities drop less quickly with age increasing (an 18 year-old can indeed still be qualified as a "young kid"). In sum, when the data is available, this method delivers models which fit humans' intuitions about the relation between numerical measure and adjective, and can handle pragmatic inference.

If we restrict attention to the 66 examples on which the Turkers completely agreed about which of these three categories was intended (again pooling 'probable' and 'definite'), then the percentage of correct inferences rises to 89% (59 correct inferences).

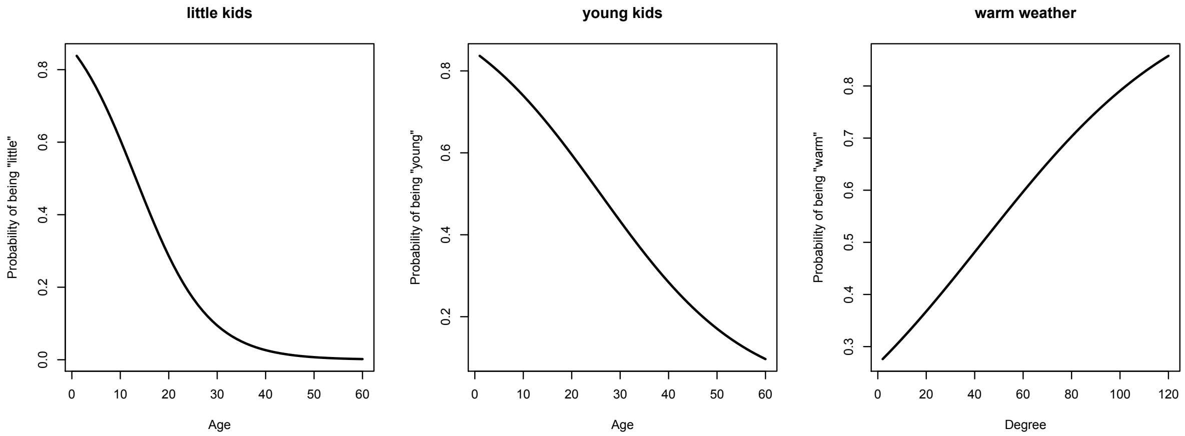


Figure 3.6: Probabilities of being appropriately described as “little”, “young” or “warm”, fitted on data retrieved when querying the Web for “little kids”, “young kids” and “warm weather”.

Figure 3.7 plots the relationship between the response entropy and the accuracy of the decision procedure I follow, along with a fitted logistic regression model using response entropy to predict whether our system’s inference was correct. The handful of empirical points in the lower left of the figure show cases of high agreement between Turkers but incorrect inference from the system. The few points in the upper right indicate low agreement between Turkers and correct inference from the system. Three of the high-agreement/incorrect-inference cases involve the adjectives *right–correct*. For low-agreement/correct-inference, the disparity could be traced to context dependency: the ordering is clear in the context of product reviews, but unclear in a television interview. The analysis suggests that overall agreement is positively correlated with the system’s chances of making a correct inference: the system’s accuracy drops as human agreement levels drop.

3.9 Discussion

This chapter focuses on the pragmatic inference needed in cases of underspecification in conversations. The goal is to build a computational model capable of recovering the kind of information that is left implicit in conversations but that participants do

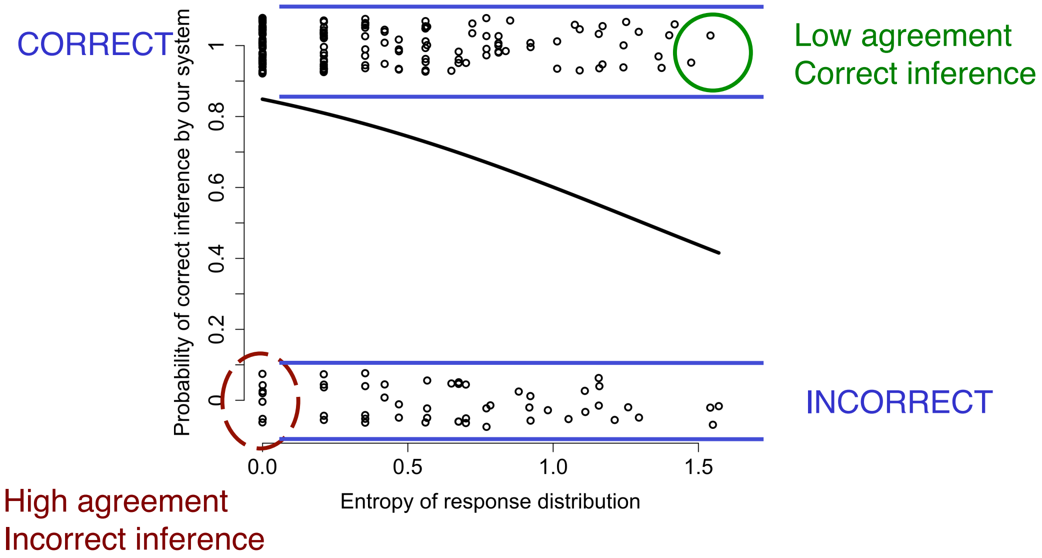


Figure 3.7: Correlation between agreement among Turkers and whether the system gets the correct answer. For each dialogue, I plot a circle at Turker response entropy and either 1 = correct inference or 0 = incorrect inference, except the points are jittered a little vertically to show where the mass of data lies. As the entropy rises (i.e., as agreement levels fall), the system's inferences become less accurate. The fitted logistic regression model (black line) has a statistically significant coefficient for response entropy ($p < 0.001$).

retrieve. I concentrate on indirect answers to *yes/no* questions. Through a corpus study, I show that they are not uncommon in dialogue. I also emphasize the uncertainty that can be present in implicit answers and the need to adequately model such uncertainty. I suggest that probabilistic models are inherently well-suited to do this. I therefore set out to find probabilistic techniques for grounding basic meanings from text and enriching those meanings based on information from the immediate linguistic context, exploiting the large amount of situated language that is now available on the Web.

In particular, I focus on gradable modifiers, seeking to learn scalar relationships between their meanings and to obtain an empirically grounded, probabilistic understanding of the clear and fuzzy cases that they often give rise to (Kamp & Partee 1995). I show that it is possible to learn the orderings people assign to gradable modifiers using Web review corpora. The technique I propose learns orderings such as *enjoyable* < *good* < *superb* or *awful* > *bad* > *disappointing*. It can do more than only assign positive or negative polarity to adjectives: the orderings can be used to drive pragmatic inference in indirect responses.

One drawback of the technique proposed here is its dependence on the availability of data. It is especially striking in the case of answers containing a numerical measure: there are multiple examples for which the system does not have access to sufficient data to learn the strength of the relation between the adjective in the question and the numerical measure in the answer. Another limitation concerns the amount of context that is taken into account. So far the techniques rely on immediate linguistic context. It is sufficient when the inferences are systematic given limited context, as in the case of *little kids*, but not when the inferences depend on a broader context, as in the case of whether a house's surface area is considered large or not: such an interpretation is highly dependent on the location (identical square footages might be considered large in London but small in Beverly Hills).

Contrary to methods that rely on existing taxonomies, the techniques I developed offer the advantage of identifying intended relations between modifiers even if they are not in an explicit scalar relation, but rather the relation needs to be constructed via world knowledge inferencing. For example, I learn not only prototypical antonyms, but also antonyms such as *qualified/young* (35), *confident/nervous* (52), *acceptable/unprecedented* (53). The techniques proposed here thus reach the level of pragmatic meaning. One advantage of working with real data is to uncover such pairs of opposite adjectives, which might not be present in taxonomies.

In this chapter, I show how to ground the meanings of gradable adjectives in a way that can successfully drive pragmatic inferences in dialog. I also point out the importance of dealing with uncertainty. Unlike prior work, I emphasize that a non-categorical modeling of indirect answers better fits the phenomenon, and the techniques I propose are probabilistic in nature.

Similar techniques, using Web data, could help us learn about the meaning and inferences attached to multidimensional adjectives. Multidimensional adjectives, such as *healthy* or *good*, are associated with many different dimensions simultaneously (Kamp 1975; Klein 1980). An adjective like *healthy*, for example, can be associated with blood pressure, pulse, cholesterol, flu, etc. (Sassoon 2010). There is a cluster of characteristics with respect to which the adjective can be evaluated. It seems okay to utter *She has high blood pressure, but is otherwise healthy*, but probably weird to say *There is almost no pulse, but he is healthy.* What counts as being healthy or not? Similarly, what does it mean for a car or a house to be in *good condition*? With respect to which dimensions are such utterances evaluated? Can we learn this automatically? Web data certainly carries much of this information, and making use of such data should provide adequate answers.

Chapter 4

Inferring relationships between text passages: Detection of conflicting information

4.1 Introduction

One aspect of the pragmatic meaning we retrieve arises in the connections we make between different pieces of text. Not only do we understand what separate sentences or utterances mean, but we also infer relationships between them. Often such relationships require using enriched pragmatic meanings, and not just literal meanings. Reading headlines such as, “Obama hasn’t been able to generate employment” and “Obama is doing a good job”, one will immediately conclude that the pieces of news make incompatible reports, but not for logical reasons. To regard these as incompatible, we need to know about the U.S. employment situation, its economy, and the link between the creation of jobs and the economy: if Obama did not create new employment, it is not helping the U.S. economy, and he is therefore not being successful.

Two basic relationships between text passages are entailment and contradiction. Condoravdi *et al.* (2003) first emphasized the importance of handling both entailment and contradiction to provide real text understanding: “Relations of entailment and contradiction are the key data of semantics, as traditionally viewed as a branch of linguistics. The ability to recognize such semantic relations is clearly not a *sufficient* criterion for language understanding: there is more than just being able to tell that one sentence follows from another. But we would argue that it is a minimal, *necessary* criterion." (p. 38). We also draw more precise relationships between pieces of text, such as the ones defined by the rhetorical structure theory (RST, Mann & Thompson 1988) briefly introduced in the previous chapter. Here, I will concentrate on contradiction detection, which I will situate in the broader context of "recognizing textual entailment" (RTE) (Dagan *et al.* 2006).

My goal in this chapter is to analyze the nature of conflicting information which appears in naturally-occurring text, and to provide a definition of "contradiction" suitable for NLP tasks. I will argue that such a definition needs to rely on the pragmatic meaning of a text rather than on its literal meaning, and that, contrary to the prevailing view in computational semantics, the phenomenon should therefore not be restricted to a logical notion. I will also describe a system to automatically identify these "contradictions". To my knowledge, it is the first system that targets contradiction detection in a broad sense. Condoravdi *et al.* (2003) restrict entailment and contradiction to a logical definition. They use a clausal representation derived from approaches in formal semantics, but do not report any empirical results for their system. Harabagiu *et al.* (2006) give the first empirical results for contradiction detection but use constructed data and focus on specific kinds of contradiction: those featuring overt negation as well as those formed by paraphrases. Yet, contradictions are not limited to these constructions; to be practically useful, a system must aim to provide broader coverage.

4.1.1 Definition of the RTE task

In recognizing textual entailment, systems are given pairs of sentences, called *text* (T) and *hypothesis* (H). The goal is to identify whether the hypothesis follows from the text and general background knowledge, according to the intuitions of an intelligent human reader. This task is already latent within question answering (Pasca &

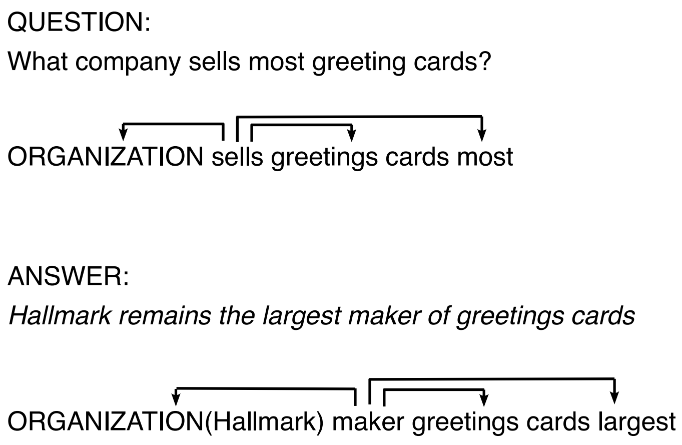

QUESTION:
What company sells most greeting cards?
ORGANIZATION sells greetings cards most
ANSWER:
Hallmark remains the largest maker of greetings cards
ORGANIZATION(Hallmark) maker greetings cards largest

Figure 4.1: Example of Question/Answer treatment (from Pasca & Harabagiu 2001).

Harabagiu 2001; Moldovan *et al.* 2003). As schematized in Figure 4.1, the question "What company sells most greetings cards?" can be viewed as a statement containing a variable (*what company*) which in this case is of the ORGANIZATION type. If the system finds a text passage that loosely entails the statement and contains a possible assignment for the variable (mainly a concept of the same type), the variable assignment is taken as the answer to the question. In this example, the passage "Hallmark remains the largest maker of greetings cards" entails the question, *Hallmark* is of type ORGANIZATION, and will be the answer to the question. This example emphasizes that the entailment is not always trivial, and might require some world knowledge. Here, as pointed out by Pasca & Harabagiu (2001), one needs to know the relationship between producing or making goods and selling them in order to retrieve "Hallmark remains the largest maker of greetings cards" as a good candidate answer to the question "What company sells most greetings cards?"

Recognizing textual entailment was formalized as a NLP task a few years ago, through the PASCAL recognizing textual entailment (RTE) challenges (Dagan *et al.* 2006; Bar-Haim *et al.* 2006; Giampiccolo *et al.* 2007) and related work within the U.S. Government AQUAINT program. From the start, the goal behind RTE has been to define as natural a task as possible. The standard is not whether the hypothesis is logically entailed, but whether it can reasonably be inferred: "We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge." (Dagan *et al.* 2006). Thus in essence, the RTE task appeals to pragmatic meaning, rather than to literal meaning.

However, this definition faced some criticisms among the members of the NLP community, and researchers argued for (Manning 2006) and against (Zaenen *et al.* 2005; Crouch *et al.* 2006) the soundness of the task. The main critique focused on restricting the task to a specific set of inference types and providing an explicit definition of background knowledge. This whole debate might have arisen from the term "entailment", which is a purely technical term in logic: a conclusion must necessarily follow from premises in every possible situation in which the premises are true for it to be entailed. However the RTE task allows *plausible inferences*. Researchers in the community therefore suggested the use of the term "inference" instead of "entailment" (Zaenen *et al.* 2005; Manning 2006) to reflect that view. This suggestion has been adopted: workshops dedicated to RTE have been called "TextInfer – Workshop on Applied Textual Inference" since 2009.

The original task definition simply draws on "common-sense" understanding of language (Chapman 2005), and focuses on how people interpret utterances. By its very nature then, the RTE task targets pragmatic meaning: of course, we might not know whether the speaker intended to convey content p with his utterance, but we can estimate how coherent p is when situated in the context of other propositions. The informal definition does not pose any problem. As pointed out by Manning (2006), people tend to agree on such natural tasks, even if they are loosely defined. Dolan *et al.* (2005) found that an informal task specification about equivalence between text passages (i.e., what counts as paraphrases) led to high annotator agreement (83%). In the case of RTE, a Mechanical Turk experiment on a subset of the RTE data by Zaenen (forthcoming) demonstrates that people agree in their judgments (85.7% agreement between the RTE annotation and Turkers' responses).

4.1.2 Definition of "contradiction detection"

In the same spirit as what was done for entailment in RTE, I propose to define contradiction detection in a pragmatic sense, relying on readers' intuitions rather than on a literal, logical sense. Whereas a logical definition of contradiction states that "sentences A and B are contradictory if there is no possible world in which A and B are both true", I will use the following definition: "two pieces of text are contradictory if they are extremely unlikely to be considered true simultaneously". Contrary to the logical definition of contradiction, the latter captures human intuitions of "incompatiblity", which suits the needs of practical NLP applications seeking to highlight discrepancies in descriptions of a same event. If *John thinks that he is incompetent*, and *his boss believes that John is not being given a chance*, one would like to detect that the targeted information in the two sentences is contradictory, even though there is a possible world in which the two sentences can be true simultaneously. Detecting such contradictions is central to many types of information analysis. It is particularly important for analysts to be made aware of conflicting factual claims and divergent viewpoints, which might reflect different sources, political leanings, or even disinformation. Consider the texts in (57):

(57) (a) Maitur Rehman, a 29-year-old Pakistani from Multan in Punjab, is reported to be the present amir of Jundullah. He had previously served in the Lashkar-e-Jhangvi, an anti-Shia terrorist organisation.

(b) Intelligence sources in the U.S. and Pakistan tell NBC News that Maitur Rehman is a low-level militant operating in South Waziristan.

If one wants to know what is the status of Maitur Rehman, it would be appropriate to retrieve both passages, and let the user know that the information found diverges: if Maitur Rehman is an amir (a leader), then he is not a low-level militant. The pieces of text contain conflicting information, and determining the exact status of Maitur Rehman would demand further analysis. I therefore mark them as contradictory.

The term "contradiction" may be subject to the same logical connotation carried by the term "entailment", and "conflicting information" might be a term better suited to my purposes, but I will use both interchangeably.

4.2 A corpus of conflicting information

To analyze how conflicting information arises and to build an automatic system targeting its identification, I manually annotated the first 3 RTE datasets for contradictions (giving rise to a total of 4,567 contradictory pairs). To do so I developed annotation guidelines (section 4.2.1), which emphasize the pragmatic nature of contradiction. What I want to detect are incompatible statements about the same event, rather than logical contradictions. In so doing I developed the first available corpus targeting contradiction defined in a broad sense, conforming to human intuitions of the phenomenon.

4.2.1 Annotation guidelines for marking contradictions in the RTE datasets

In this section, I reproduce verbatim the annotation guidelines that I developed and followed for my own data annotation. The same guidelines were used by assessors at NIST (National Institute of Standards and Technology) for annotating the RTE3 pilot experiment (Voorhees 2008), and similar guidelines were used in the RTE-4 and RTE-5 evaluations run by NIST.

Recognizing Textual Entailment (RTE) items consist of two pieces of text, a brief text (T) and a short hypothesis (H). For some, the hypothesis follows from the text (that is, a normal reader would be happy to accept the text as strong evidence that the hypothesis is true, assuming that the text is reliable). This is technically referred to as “entailment”. These items are marked “YES”. You shouldn’t change these. For the rest, we wish to distinguish between whether the text and hypothesis are contradictory, which we will label “NO”, or whether the two pieces contain overlapping or different information but the hypothesis neither follows from or contradicts the text, which we will label “UNKNOWN”.

Definition of contradiction

To decide if the text and hypothesis are contradictory, ask yourself the following question: If I were shown two contemporaneous documents one containing each of these passages, would I regard it as very unlikely that both passages could be true at the same time? If so, the two contradict each other. Another way of stating this would be: the hypothesis is contradictory if assertions in the hypothesis appear to directly refute, or show portions of the text to be false/wrong, if the hypothesis were taken as reliable. You should be able to state a clear basis for a contradiction, such as "the text says the group traveled west to Mosul, while the hypothesis says they were traveling from Syria (which is to the east of Mosul)." For example, the following are contradictions:

(58) [RTE1_test 828] contradiction

T: Jennifer Hawkins is the 21-year-old beauty queen from Australia.

H: Jennifer Hawkins is Australia's 20-year-old beauty queen.

(59) [RTE2_ dev 404] contradiction

T: In that aircraft accident, four people were killed: the pilot, who was wearing civilian clothes, and three other people who were wearing military uniforms.

H: Four people were assassinated by the pilot.

You should mark as a contradiction a text and hypothesis reporting contradictory statements, if the reports are stated as facts. We can see these as carrying an embedded contradiction. For example:

(60) [RTE2_dev 320] contradiction

T: That police statement reinforced published reports, that eyewitnesses said de Menezes had jumped over the turnstile at Stockwell subway station and was wearing a padded jacket, despite warm weather.

H: However, the documents leaked to ITV News suggest that Menezes, an electrician, walked casually into the subway station and was wearing a light denim jacket.

For something to be a contradiction, it does not have to be impossible for the two reports to be reconcilable, it just has to appear highly unlikely in the absence of further evidence. For instance, it is reasonable to regard the first pair below as a contradiction (it is not very plausible that the bodies – of someone who has a secretary, etc. – were not found for over 18 months), but it does not seem prudent to regard the second pair as contradictory (despite a certain similarity in the reports, they could easily both be true):

(61) [RTE1_dev 579] contradiction

T: The anti-terrorist court found two men guilty of murdering Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991.

H: Shapour Bakhtiar died in 1989.

(62) [RTE1_test 2113] unknown (not a contradiction)

T: Five people were killed in another suicide bomb blast at a police station in the northern city of Mosul.

H: Five people were killed and 20 others wounded in a car bomb explosion outside an Iraqi police station south of Baghdad.

How to interpret the data?

I. Noun phrase coreference:

Compatible noun phrases between the text and the hypothesis should be treated as coreferent in the absence of clear countervailing evidence. For example, below we should assume that the two references to *a woman* refer to the same woman:

(63) [RTE1_dev 201] contradiction

T: Passions surrounding Germany's final match at the Euro 2004 soccer championships turned violent when a woman stabbed her partner in the head because she didn't want to watch the game on television.

H: A woman passionately wanted to watch the soccer championship.

Similarly, references to dates like “Thursday” should be assumed to be coreferent in the absence of countervailing evidence.

II. Event coreference:
Whether to regard a text and hypothesis as describing the same event is more subtle. If two descriptions appear overlapping, rather than completely unrelated, by default assume that the two passages describe the same context, and contradiction is evaluated on this basis. For example, if there are details that seem to make it clear that the same event is being described, but one passage says it happened in 1985 and the other 1987, or one passage says two people met in Greece, and the other in Italy, then you should regard the two as a contradiction. Below, it seems reasonable to regard “a ferry collision” and “a ferry sinking” as the same event. The reports then make contradictory claims on casualties:

(64) [RTE2_dev 237] contradiction

T: Rescuers searched rough seas off the capital yesterday for survivors of a ferry collision that claimed at least 28 lives, as officials blamed crew incompetence for the accident.

H: 100 or more people lost their lives in a ferry sinking.

In other circumstances, it is most reasonable to regard the two passages as describing different events. You have to make your best judgment, given the limited information available. You should use world knowledge about the frequency of event types in making this decision. For instance, example (62) above was not marked as a contradiction, as it does not seem compelling to regard “another suicide bomb blast” and “a car bomb explosion” as referring to the same event. And for the two passages below, there just doesn’t seem much evidence that they have anything to do with each other:

(65) [RTE2_dev 333] unknown (not a contradiction)

T: The European-born groups with the highest labor force participation rates were from Bosnia and Herzegovina.

H: The European country with the highest birth rate is Bosnia-Herzegovina.

In the general RTE guidelines, it says the text and the hypothesis are meant to be regarded as roughly contemporaneous, but may differ in date by a few days, and so details of tense are meant to be ignored when deciding whether a text entails the hypothesis or not. However in an example like the following, it seems clear that the hypothesis is not possible as a consistent, contemporaneous statement with the text, and so we mark it as contradictory:

(66) [RTE3_dev 357] contradiction

T: The Italian parliament may approve a draft law allowing descendants of the exiled royal family to return home. The family was banished after the Second World War because of the King's collusion with the fascist regime, but moves were introduced this year to allow their return.

H: Italian royal family returns home.

4.2.2 Annotation results

Using the above annotation guidelines, I found that contradictions constitute approximately 10% of the first 3 RTE datasets. Table 4.1 gives the number of contradictions in each dataset. The RTE datasets are balanced between entailments and non-entailments, and even in these datasets targeting inference, there are few contradictions.

The RTE datasets were based on real data and NLP tasks, but examples were nevertheless chosen manually to target textual inference, and the text was also tampered with to obtain the same number of entailments and non-entailments. Thus, they might not reflect "real-life" contradictions. I therefore also collected contradictions "in the wild", from different datasets in which I can assume that the text is not

Data	# contradictions	# total pairs
RTE1_dev1	48	287
RTE1_dev2	55	280
RTE1_test	149	800
RTE2_dev	111	800
RTE2_test	104	800
RTE3_dev	80	800
RTE3_test	72	800

Table 4.1: Number of contradictions in the first three RTE datasets.

altered. The resulting corpus contains 131 contradictory pairs: 19 from newswire, mainly from looking at related articles in Google News, 51 from Wikipedia searching for contradictory tags in the history, 10 from the Lexis Nexis database, and 51 from the data prepared by LDC for the distillation task of the DARPA GALE program,[1] which used real data. Despite the fact that this collection draws on several, very different sources of text, I argue that this corpus better reflects naturally-occurring contradictions than the RTE datasets.[2]

As for RTE, the vagueness of the definition of contradiction does not hurt inter-annotator agreement. As mentioned above, I provided my annotation guidelines to NIST for their RTE3 pilot experiment (Voorhees 2008) in which systems were asked to mark pairs of sentences as entailed, contradictory, or neither.[3] One of the datasets, the RTE3_test set, has been independently annotated by NIST: I found a high inter-annotator agreement ($\kappa = 0.81$) between their annotations and mine (Cohen 1960), showing that, even when limited context is available, humans tend to agree on what a contradiction is.[4]

[1] The goal of the DARPA GALE program is to develop and apply computer software technologies to analyze and interpret huge volumes of speech and text in multiple languages. The distillation task aims at identifying information relevant to a user's query, and delivering it in easy-to-understand forms.

[2] The corpus is available at http://nlp.stanford.edu/projects/contradiction.

[3] Further information about this task can be found at http://nlp.stanford.edu/RTE3-pilot/.

[4] This contrasts with the low inter-annotator agreement reported by Sanchez-Graillet & Poesio (2007) for contradictions in descriptions of protein-protein interactions. The only hypothesis I have to explain this contrast is the difficult readability of scientific material.

4.3 Typology of conflicting information

There are different ways in which conflicting information arises in text. Information conflicts appear when facts diverge: for example, if detached from a time line, articles can show contradictory figures for a rising death toll as in (64) above.

Conflicting opinions and reports are also very frequent in texts: one example is the pair of sentences in (67) repeated from (57) ; another example is given in (68) repeated from (60) where two different sources report conflicting data.

(67) (a) Maitur Rehman, a 29-year-old Pakistani from Multan in Punjab, is reported to be the present amir of Jundullah. He had previously served in the Lashkar-e-Jhangvi, an anti-Shia terrorist organisation.

(b) Intelligence sources in the U.S. and Pakistan tell NBC News that Maitur Rehman is a low-level militant operating in South Waziristan.

(68) [RTE2_dev 320]

T: That police statement reinforced published reports, that eyewitnesses said de Menezes had jumped over the turnstile at Stockwell subway station and was wearing a padded jacket, despite warm weather.

H: However, the documents leaked to ITV News suggest that Menezes, an electrician, walked casually into the subway station and was wearing a light denim jacket.

Even though logically the two sentences can both be true at the same time, one would like to detect that the embedded information about *de Menezes* in the two texts in (68) differs.

Indicating conflicting views can also be done in very subtle ways, by providing different perspectives on the same event. I will only illustrate this with one instance: cross-examination in a courtroom. In this case, the sequential position of the terms, rather than the terms themselves, gives rise to conflicting statements (Drew 1992). The examples come from part of a rape trial, in which the victim was being cross-examined by the defendant's attorney.

(69) A: Well you had some fairly lengthy conversations with the defendant, didn't you?

A: On that evening of February fourteenth?

B: Well, we were all talking.

As noted by Drew, "the terms *all talking* and *lengthy conversations* are not by themselves incompatible" (p. 491). Nevertheless with her reply "we were all talking", the witness is challenging the attorney's characterization of the scene and proposes an alternative version of it. This is also reminiscent of the examples of indirect answers from the previous chapter: B is not directly answering the attorney's question, leaving room for inference. The same phenomenon appears in the description of the victim and defendant's placement in the bar. The attorney uses the terms "sitting with you", to which the witness replies with the terms "sitting at the same table". Again Drew notes: "In a conversational setting it may be doubted that in describing someone joining one at one's table for a drink, there is sufficient difference between that person *sitting with one* and *sitting at one's table* for it to be worth troubling to insist on the latter version. [...] However, in not allowing the attorney's versions to pass unamended, the witness orients to the differences between these versions for her story." (p. 492). These last two examples emphasize the existence of different levels of (in)compatibility: the descriptions of the event are not incompatible but there is an issue of compatibility in the *perspective* actors have on the event.

In my work on contradiction detection, I will only be concerned with incompatible descriptions of an event, excluding perspective shifts on an event. I focus thus on divergent facts and conflicting reports. Contradictions may arise in a number of different constructions, some overt and others that are complex to detect. I examined the contradictory pairs in the RTE datasets that I annotated as well as the "real-life" contradictions that I gathered. I make a distinction between two primary categories of contradiction: (I) those occurring via antonymy, negation, and date/number mismatch, and (II) contradictions arising from the use of factive or modal words, structural and subtle lexical contrasts, as well as world knowledge.

Category I contradictions are more often overt contradictions, which are relatively easy to detect, as in the death toll example (64) or in the *de Menezes* example (60). Additionally, little external information is needed to gain broad coverage of antonymy, negation, and numeric mismatch contradictions; each involves only a closed set of words or data that can be obtained using existing resources and techniques (e.g., WordNet (Fellbaum 1998), VerbOcean (Chklovski & Pantel 2004)), and no domain knowledge is necessary.

The contradictions in the second category are more difficult to find automatically: they involve lexical and structural discrepancies, as well as inconsistency via world knowledge. Such contradictions require complex and precise models of sentence meaning. In (57) for instance, the meaning of *amir* is crucial for detecting the contradiction. To find the contradiction in (70), it is necessary to learn that *X said Y did nothing wrong* and *X accuses Y* are incompatible.

(70) [RTE3_dev 160]

T: The Canadian parliament's Ethics Commission said former immigration minister, Judy Sgro, did nothing wrong and her staff had put her into a conflict of interest.

H: The Canadian parliament's Ethics Commission accuses Judy Sgro.

Presently, there exist methods for learning loosely opposing terms (Marcu & Echihabi 2002) and paraphrase learning has been thoroughly studied, but successfully extending these techniques to learn incompatible phrases poses difficulties because of the data distribution. Lexical complexities and variations in the function of arguments across verbs can make recognizing structural contradictions complicated. Even when similar verbs are used and clear argument differences exist, structural differences may indicate either non-entailment or contradiction, and distinguishing the two automatically is problematic. Consider the contradictory pair (71) and the pair (72) which is not a contradiction:

(71) [RTE2_test 37]

T: The Channel Tunnel stretches from Cheriton, Kent in England to the town of Sangatte in the Nord Pas-de-Calais region of France. It is the second-longest rail tunnel in the world, the longest being a tunnel in Japan.

H: The Channel Tunnel connects France and Japan.

(72) [RTE2_test 401]

T: The CFAP purchases food stamps from the federal government and distributes them to eligible recipients.

H: A government purchases food.

In both cases, the first sentence discusses one entity (*The Channel Tunnel*, *CFAP*) which has a relationship (*stretch*, *purchase*) to other entities. The second sentence (b) posits a similar relationship that includes one of the entities involved in the original relationship as well as an entity that was not involved. However, different outcomes result because a tunnel can only connect two unique locations whereas more than one entity may purchase food. These frequent interactions between world knowledge and structure make it hard to ensure that any particular instance of structural mismatch is a contradiction. Further, in (73), one needs to have some knowledge about head companies and branches to detect the incompatibility.

(73) [RTE1_dev 2084]

T: Microsoft Israel, one of the first branches outside the USA, was founded in 1989.

H: Microsoft was established in 1989.

Table 4.2 gives the distribution of contradiction types for RTE3_dev, RTE3_test and the real contradiction corpus. Globally, we see that contradictions in category II occur frequently and dominate the RTE datasets. In the real contradiction corpus, there is a much higher rate of type I contradictions (negation and numeric). Lexical contradictions are also very frequent in the real contradiction corpus. This supports the intuition that in the real world, contradictions primarily occur for two reasons: information is updated as knowledge of an event is acquired over time (e.g., a rising death toll) or various parties have divergent views of an event.

	Type	RTE3_dev	RTE3_test	'Real' corpus
I	Antonym	15.0	9.7	9.2
	Negation	8.8	6.9	17.6
	Numeric	8.8	9.7	29.0
II	Factive/Modal	5.0	13.9	6.9
	Structure	16.3	26.4	3.1
	Lexical	18.8	16.7	21.4
	World-knowledge	27.5	16.7	13.0

Table 4.2: Percentages of contradiction types in the RTE3_dev dataset, the RTE3_test dataset and the real contradiction corpus.

4.4 System description

The system I developed to detect contradiction is an adaption of the Stanford RTE system (MacCartney *et al.* 2006). Given pairs of passages called text (T) and hypothesis (H), it decides whether or not they are contradictory. The system follows the Stanford system's multi-stage architecture.

The first stage computes the linguistic representations containing information about the semantic content of the passages: the text and hypothesis are converted to typed dependency graphs produced by the Stanford parser (Klein & Manning 2003; de Marneffe *et al.* 2006). To improve the dependency graph as a pseudo-semantic representation, collocations in WordNet and named entities are collapsed, causing entities and multiword relations to become single nodes.

The second stage provides an alignment between the graphs, consisting of a mapping from each node in the hypothesis to a unique node in the text or to null. Each alignment is assigned a score, which tries to capture how well the two passages could be aligned. The scoring measure uses node similarity (irrespective of polarity) and structural information based on the dependency graphs. Similarity measures and structural information are combined via weights learned using the passive-aggressive online learning algorithm MIRA (Crammer & Singer 2001). Alignment weights were learned using manually annotated RTE development sets (see Chambers *et al.* (2007)).

Figure 4.2 gives an example of dependency representation and graph alignment between the text/hypothesis pair in (74). Further details about the scoring alignment measure and the search algorithm can be found in (de Marneffe *et al.* 2007; Chambers *et al.* 2007).

(74) T: CNN reported that several troops were killed in today's ambush.

H: Thirteen soldiers lost their lives in the ambush.

In the final stage, a set of features[5] targeting the detection of contradictions are extracted, to which a logistic regression is applied to classify the pair as contradictory or not. Feature weights are hand-set, guided by linguistic intuition. The features rely on mismatches between the text and the hypothesis. However pairs of sentences which do not describe the same event, and thus cannot contradict one another, could nonetheless contain mismatching information. An extra stage to filter non-coreferent events is therefore added before feature extraction. For example, in (75), it is necessary to recognize that *the Johnstown Flood* has nothing to do with *a ferry sinking*; otherwise conflicting death tolls result in labeling the pair a contradiction.

(75) [RTE2_dev 208]

T: More than 2,000 people lost their lives in the devastating Johnstown Flood.

H: 100 or more people lost their lives in a ferry sinking.

This issue does not arise for textual inference: elements in the hypothesis not supported by the text lead to non-inference, regardless of whether the same event is described. For contradiction, however, it is critical to filter unrelated sentences to avoid finding false evidence of contradiction when there is contrasting information about different events.

[5] In this dissertation, I am using the term *features* with its machine learning sense, and not in the way it is used in various areas of linguistics. It refers to an elementary pattern weighed in a machine learning classifier.

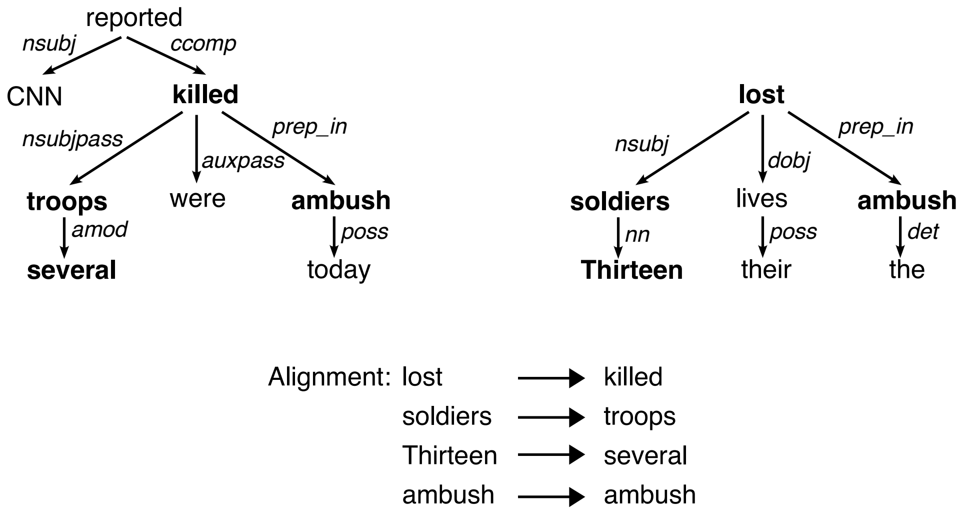


Figure 4.2: Dependency graphs of text T *CCN reported that several troops were killed in today's ambush* and hypothesis H *Thirteen soldiers lost their lives in the ambush*, as well as alignment from hypothesis to text.

4.4.1 Filtering non-coreferent events

This stage aims at removing pairs of sentences which do not describe the same event, and can therefore not be considered in a relation of contradiction to one another. If (76) is not identified as describing non-coreferent events, the conflicting diameters would be seen as an indication of contradiction, and the pair would incorrectly be tagged as contradictory.

(76) T: Pluto's moon, which is only about 25 miles in diameter, was photographed 13 years ago.

H: The moon Titan has a diameter of 5100 kms.

To identify pairs of sentences describing non-coreferent events, the system uses a crude filter based on the peculiarities of the RTE data and on topicality. Given the structure of RTE data, in which the hypotheses are shorter and simpler than the texts, one straightforward strategy for detecting coreferent events is to check whether the root of the graph representing the hypothesis sentence is aligned in the graph representing the text sentence. In (74) for example, the root of the hypothesis sentence *lost* is aligned to a word in the text sentence (see figure 4.2). If the root of the hypothesis cannot be aligned to a word in the text sentence, it probably means that the hypothesis sentence is not related to the topic of the text sentence. However, some RTE hypotheses are testing systems' abilities to detect relations between entities (e.g., *John of IBM ...* → *John works for IBM*; *John from Sunnyvale, CA ...* → *John lives in Sunnyvale*, etc.). Thus, for verbs indicating such relations (e.g., *work*, *live*), even if they were unaligned roots in the hypothesis, I did not take that as an indication of event non-coreference. As shown in table 4.3, checking the root alignment improves results on RTE data. However, the assumption of directionality made in this strategy does not carry over to real world data, and we cannot assume that one sentence will be short and the other more complex.

Assuming two sentences of comparable complexity, I hypothesize that topicality could be a good indicator of whether the sentences describe the same event. The text and the hypothesis each receive a topicality score, and the text and hypothesis are considered topically related if either topicality score is above a threshold tuned on the training sets. The topicality score is calculated as follows. There is a continuum of topicality from the start to the end of a sentence (Firbas 1971). I thus originally defined the topicality of an NP by n^w where n is the nth NP in the sentence. Additionally, I accounted for multiple clauses in a sentence by weighting each clause equally. In (77), *Australia* receives the same weight as *Prime Minister* because each begins a clause.

(77) [RTE2_test 122]

T: Prime Minister John Howard says he will not be swayed by a videotaped warning that Australia faces more terrorism attacks unless it withdraws its troops from Iraq and Afghanistan.

H: Australia withdraws from Iraq.

However, accounting for multiple clauses by weighting each one equally did not improve results on the development set, and I thus use a simpler, unweighted model.

Strategy	Precision	Recall
No filter	55.10	32.93
Root	61.36	32.93
Root + topic	61.90	31.71

Table 4.3: Precision and recall for contradiction detection on RTE3_dev using different filtering strategies.

The topicality score of a sentence is calculated as a normalized score across all aligned NPs.[6] While modeling topicality provides an additional improvement in precision (table 4.3), some examples of non-coreferent events are still not filtered, such as (78).

(78) [RTE2_dev 29]

T: Also Friday, five Iraqi soldiers were killed and nine wounded in a bombing, targeting their convoy near Beiji, 150 miles north of Baghdad.

H: Three Iraqi soldiers also died Saturday when their convoy was attacked by gunmen near Adhaim.

It seems that the nature of the nominals involved in the event description needs to be taken into account. Compare (78), which contains indefinite plurals, with the following example involving a proper noun:

(79) T: Princess Diana died in Paris.

H: The car accident killing Princess Diana occurred in London.

There is a high probability in discourse that two identical proper nouns corefer: the two sentences in (79) indeed refer to a unique event, and the location mismatch renders them incompatible, whereas in (78) the attacks are likely to refer to different events and the two sentences do not present mismatching information (i.e., different location) about the same event.

[6] Since dates can often be viewed as scene setting rather than what the sentence is about, I ignore these in the model. However, ignoring or including dates in the model creates no significant differences in performance on the RTE data.

4.4.2 Contradiction features

Mismatching information between sentences is often a good cue of non-entailment (Vanderwende *et al.* 2006), but it is not sufficient for detecting contradiction which requires more precise comprehension of the consequences of sentences. Some of the features used in the Stanford RTE system have been more precisely defined to only capture mismatches in similar contexts, instead of global mismatches. These features are described below.

Antonymy features. Antonyms which are aligned between the text and the hypothesis are a very good cue to contradiction. The list of antonyms and contrasting words comes from WordNet, from which I extract words with direct antonymy links and expand the list by adding words from the same synset as the antonyms. I also use loosely opposing verbs from VerbOcean (Chklovski & Pantel 2004). I check whether an aligned pair of words appears in the list, as well as checking for common negative prefixes (e.g., *anti-*, *un-*). The polarity of the context is used to determine if the antonyms create a contradiction.

Polarity features. Polarity difference between the text and hypothesis is often a good indicator of contradiction, provided there is a good alignment:

(80) [RTE1_dev 227]

T: A closely divided U.S. Supreme Court said on Thursday its 2002 ruling that juries and not judges must impose a death sentence applies only to future cases, a decision that may affect more than 100 death row inmates.

H: The Supreme Court decided that only judges can impose the death sentence.

The polarity features capture the presence (or absence) of linguistic markers of negative polarity contexts. These markers are scoped such that words are considered negated if they have a negation dependency in the graph or are an explicit linguistic marker of negation (e.g., simple negation (*not*), downward-monotone quantifiers (*no*, *few*), or restricting prepositions). If one word is negated and the other is not, we may have a polarity difference. This difference is confirmed by checking that the words are not antonyms and that they lack unaligned prepositions or other context that suggests they do not refer to the same thing. In some cases, negations are propagated onto the governor, which allows one to see that "no bullet penetrated" and "a bullet did not penetrate" have the same polarity.

Number, date and time features. Numeric mismatches can indicate contradiction (see examples (58) and (64) above). The numeric features recognize (mis-)matches between numbers, dates, and times. Date and time expressions are normalized, and numbers are represented as ranges. This includes expression matching (e.g., *over 100* and *200* is not a mismatch). Aligned numbers are marked as mismatches when they are incompatible and the other surrounding words match well, indicating the numbers refer to the same entity.

Structural features. These features aim to determine whether the syntactic structures of the text and hypothesis create contradictory statements. For example, the system compares the subjects and objects for aligned verbs (that are not contradictory). If the subject in the text overlaps with the object in the hypothesis, we find evidence for a contradiction. Consider:

(81) [RTE2_dev 84]

T: Jacques Santer succeeded Jacques Delors as president of the European Commission in 1995, the second Luxemburger to hold this high office.

H: Delors succeeded Santer in the presidency of the European Commission.

In the text, the subject of *succeed* is *Jacques Santer* while in the hypothesis, *Santer* is the object of *succeed*, suggesting that the two sentences are incompatible.

Modality features. Simple patterns of modal reasoning are captured by mapping the text and hypothesis to one of six modalities ((*not_*)*possible*, (*not_*)*actual*, (*not_*)*necessary*), according to the presence of predefined modality markers such as *can* or *maybe*. A feature is produced if the text/hypothesis modality pair gives rise to a contradiction. For instance, the following pair will be mapped to the contradiction judgment (*possible*, *not_possible*):

(82) [RTE1_dev 524]

T: The trial court may allow the prevailing party reasonable attorney fees as part of costs.

H: The prevailing party may not recover attorney fees.

However as pointed out by Manning (2006), *may* or *can* are also used as a form of hedging, especially in scientific or political discourse. He gives as examples the following pairs from the RTE datasets (83) and (84) which are marked as entailments.

(83) [RTE1_dev 19]

T: Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake – this kind of consumption apparently also can help reduce the risk of diseases.

H: Coffee drinking has health benefits.

(84) [RTE1_dev 20]

T: Eating lots of foods that are a good source of fiber may keep your blood glucose from rising too fast after you eat.

H: Fiber improves blood sugar control.

Hedging can appear with contradictory texts too, but the system cannot handle these nuances yet. I therefore chose not to classify the modality pair (*possible*, *actual*) as a marker of contradiction. A pair such as (85) would thus not be correctly captured as contradictory.

(85) T: Suncreams designed for children could offer less protection than they claim on the bottle.

H: Suncreams designed for children protect at the level they advertise.

Factivity features. The context in which a verb phrase is embedded may give rise to contradiction:

(86) [RTE1_test 1981]

> T: The bombers had not managed to enter the embassy compounds.
>
> H: The bombers entered the embassy compounds.

Negation influences some factivity patterns: *Bill forgot to take his wallet* contradicts *Bill took his wallet* while *Bill did not forget to take his wallet* does not contradict *Bill took his wallet*. For each text/hypothesis pair, the system checks the (grand)parent of the text word aligned to the hypothesis verb, and generates a feature based on its factivity class. In the example above, *entered* will be aligned with the verb *enter* in the text T. The system will check the parent of *enter*, in this case *managed*, as well as negation (the parent is negated but the verb in H is not negated). The rule for the implicative verb *manage* will generate a feature indicating that in this case there is contradiction. Factivity classes are formed by clustering my expansion of the PARC lists of factive, implicative and non-factive verbs (Nairn *et al.* 2006) according to how they create contradiction.

Relational features. A large proportion of the RTE data is derived from information extraction tasks where the hypothesis captures a relation between elements in the text. Using Semgrex, a pattern matching language for dependency graphs, the system finds such relations and ensures that the arguments between the text and the hypothesis match. In (87), it detects that *Fernandez* works for *FEMA*, and that because of the negation, a contradiction arises.

(87) [RTE2_dev 207]

> T: Fernandez, of FEMA, was on scene when Martin arrived at a FEMA base camp before going to the hospital.
>
> H: Fernandez doesn't work for FEMA.

Relational features provide accurate information but are difficult to extend for broad coverage.

4.5 Evaluation and results

The contradiction detection system presented here was developed using all datasets listed in the first part of table 4.4. As test sets, I used RTE1_test, the independently annotated RTE3_test, and Neg_test (which I detail below). I focused on attaining high precision. In a real world setting, it is likely that the contradiction rate is quite low; rather than overwhelming true positives with false positives, rendering the system impractical, it marks contradictions conservatively focusing on attaining high precision.

As mentioned earlier, Harabagiu *et al.* (2006) presented the main previous evaluation of detecting contradictions. Unlike their work, I focus on detecting contradictions appearing in any type of construction. Harabagiu *et al.* (2006) concentrated on contradictions featuring overt negation as well as on contradictions arising in paraphrases. They constructed two corpora on which they evaluated their system. One (LCC_negation) was created by overtly negating each entailment in the RTE2 data, producing a balanced dataset. To avoid overtraining, negative markers were also added to each instance of non-entailment while ensuring that these markers did not create contradictions. Their second corpus (LCC_paraphrase) was produced by paraphrasing the hypothesis sentences from LCC_negation, in a way that did not contain a negative element such as *not* or *never*: “A hunger strike was not attempted” was loosely paraphrased by “A hunger strike was called off”. They achieved very good performance: accuracies of 75.63% on LCC_negation and 62.55% on LCC_paraphrase.

Because their corpora are constructed using negation and paraphrase, they are unlikely to cover all types of contradictions mentioned in section 4.3. We might hypothesize that explicit negations are commonly eliminated through the substitution of antonyms. Imagine for instance:

(88) H: Bill has finished his math.

Neg-H: Bill hasn’t finished his math.

Para-Neg-H: Bill is still working on his math.

The rewriting in both the negated and the paraphrased corpora is likely to leave one in the space of ‘easy’ contradictions and addresses fewer than 30% of contradictions

	Precision	Recall	Accuracy
RTE1_dev1	70.37	40.43	–
RTE1_dev2	72.41	38.18	–
RTE2_dev	64.00	28.83	–
RTE3_dev	61.90	31.71	–
Neg_dev	74.07	78.43	75.49
Neg_test	62.97	62.50	62.74
LCC_negation	–	–	75.63
RTE1_test	42.22	26.21	–
RTE3_test	22.95	19.44	–
Avg. RTE3_test	10.72	11.69	

Table 4.4: Precision and recall figures for contradiction detection. Accuracy is given for balanced datasets only. 'LCC_negation' refers to performance of Harabagiu *et al.* (2006). 'Avg. RTE3_test' refers to mean performance of the 12 submissions to the RTE3 Pilot.

(see table 4.2). I contacted the LCC authors to obtain their datasets, but they were unable to make them available to me. Thus, to provide some comparison, I simulated the LCC negation corpus, adding negative markers to the RTE2_test data (Neg_test), and to a development set (Neg_dev) constructed by randomly sampling 50 pairs of entailments and 50 pairs of non-entailments from the RTE2 development set.

Table 4.4 gives the precision and recall figures for contradiction detection on the training and test sets. The results on the test sets show that performance drops on new data, highlighting the difficulty in generalizing from a small corpus of positive contradiction examples, as well as underlining the complexity of building a broad coverage system. This drop in accuracy on the test sets is greater than that of many RTE systems, suggesting that generalizing for contradiction is more difficult than for entailment. Particularly when addressing contradictions that require lexical and world knowledge, I am only able to add coverage in a piecemeal fashion, resulting in improved performance on the development sets but only small gains for the test sets. Thus, as shown in table 4.5, the systems achieves 13.3% recall on lexical contradictions in RTE3_dev but is unable to identify any such contradictions in RTE3_test.

	Type	RTE3_dev		RTE3_test	
I	Antonym	25.0	(3/12)	42.9	(3/7)
	Negation	71.4	(5/7)	60.0	(3/5)
	Numeric	71.4	(5/7)	28.6	(2/7)
II	Factive/Modal	25.0	(1/4)	10.0	(1/10)
	Structure	46.2	(6/13)	21.1	(4/19)
	Lexical	13.3	(2/15)	0.0	(0/12)
	World-knowledge	18.2	(4/22)	8.3	(1/12)

Table 4.5: Recall by contradiction type.

Additionally, the precision of category II features is less than that of category I features. Structural features, for example, are responsible for the 36 non-contradictions tagged as contradictions in RTE3_test, over 75% of the precision errors.

4.6 Error analysis

A significant issue in contradiction detection is the lack of feature generalization. This problem is especially apparent for items in category II requiring lexical and world knowledge, which proved to be the most difficult contradictions to detect on a large scale. While the system is able to find certain specific relationships in the development sets, features targeting category II contradictions attained only limited coverage. Many contradictions in this category require multiple inferences and remain beyond the system's capabilities:

(89) [RTE3_dev 156]

T: The Auburn High School Athletic Hall of Fame recently introduced its Class of 2005 which includes 10 members.

H: The Auburn High School Athletic Hall of Fame has ten members.

Of the types of contradictions in category II, the system is best at addressing those formed via structural differences and factive/modal constructions as shown in table

4.5. However, creating features with sufficient precision is an issue for these types of contradictions. Intuitively, two sentences that have aligned verbs with the same subject and different objects (or vice versa) are contradictory. This indeed indicates a contradiction 55% of the time on the development sets, but this is not high enough precision given the rarity of contradictions.

Another type of contradiction where precision falters is numeric mismatch. A high recall is obtained for this type (table 4.5), as it is relatively simple to determine if two numbers are compatible, but high precision is difficult to achieve due to differences in what numbers may mean. Consider:

(90) [RTE2_dev 565]

T: Nike Inc. said that its first-quarter profit grew 32 percent, as the world's largest sneaker and athletic apparel company posted broad gains in sales and orders.

H: Nike said orders for footwear and apparel for delivery totaled $4.9 billion, including a 12 percent increase in U.S. orders.

The system detects a mismatch between *32 percent* and *12 percent*, ignoring the fact that one refers to *profit* and the other to *orders.* Accounting for context requires extensive text comprehension; it is not enough to simply look at whether the two numbers are headed by similar words (*grew* and *increase*). This emphasizes the fact that mismatching information is not sufficient to indicate contradiction.

The system handles single word antonymy with high precision (78.9%), as well as negation. Nevertheless, Harabagiu *et al.* (2006)'s performance on detecting contradictions arising from negation and antonyms (64% on a balanced dataset) demonstrates that further improvement on these types is possible; indeed, they use more sophisticated techniques to extract opposing terms and detect polarity differences. Thus, detecting category I contradictions is feasible with current systems. Since more than half of the examples found in the real corpus belong to that category, it suggests that we may be able to gain sufficient traction on contradiction detection for real world applications. Even so, category II contradictions must be targeted to detect many of the most interesting examples and to solve the entire problem of contradiction detection. Some types of these contradictions, such as lexical and world knowledge, are currently beyond the system's grasp, but progress can be made on the structure and factive/modal types.

Another issue is the quality of the alignment between the hypothesis and text graphs. In some cases, a bad alignment is the cause of recall or precision errors, and not the contradiction detection mechanism per se:

(91) [RTE2_test 107]

T: Since Concorde's first flight in 1969, it was recognized as the safest airplane in the history of aviation. And in spite of this dramatic crash on July 25, it still remains the safest way to fly.

H: Concorde's first crash was in 1969.

In this pair, *crash* in the hypothesis is aligned to *flight* in the text, which are identified as oppositional terms, and the system therefore incorrectly marks the pair as contradictory. Finding the optimal alignment between the hypothesis and the text is a difficult search process: improvement in the alignment stage would clearly increase the performance of the system. However I leave that issue aside.

4.7 Discussion

The work in this chapter raised awareness about the importance of contradiction detection in the NLP community. As already mentioned, the chapter is based on de Marneffe *et al.* (2008). Since the publication of that paper, there has been some further progress on contradiction detection.

The results of the contradiction detection system presented here highlight the difficulty of automatically finding conflicting information on a broad scale. I am not aware of further work attempting to detect the wide spectrum of naturally-occurring contradictions. Ritter *et al.* (2008) restrict their attention to a specific category of contradictions. They investigate the automatic detection of contradictions in functional relations (e.g., *be born*) and emphasize the fact that to successfully identify incompatible functional relations, background knowledge is often necessary. Compare (92) and (93):

(92) (a) Mozart was born in Vienna.
(b) Mozart was born in Salzburg.

(93) (a) Alan Turing was born in London.
(b) Alan Turing was born in England.

To correctly identify that (92) presents contradictory information whereas (93) does not, one needs to know that Vienna and Salzburg are two different cities but that London is a city of England. Ritter *et al.* (2008) gathered a large corpus of potential contradictions involving functional relations (such as (92) and (93)) from the Web, using distributional information of the relation's arguments. Out of the 8,884 potential contradictions retrieved, only 110 were incompatible. For this unbalanced corpus, their contradiction detection system, AuContraire, achieves a precision of 62% and a recall of 19%. The system mainly uses three features to filter out non-contradictory pairs: synonyms (*died from renal failure* ∼ *died from kidney failure*), meronyms (*born in Salzburg* does not contradict *born in Austria*) and argument structure using named entities matching (*born in Salzburg* does not contradict *born in 1756*). Despite the restriction on the structure in which the contradictions appear, these results highlight again the fact that contradiction detection is an intrinsically difficult task.

A key component in contradiction detection is the identification of coreferent events. The importance of event coreference was actually recognized in the MUC information tasks (in which it was key to identify scenarios related to the same event) at the end of the eighties. But, coreferent event identification has been neglected lately in the NLP community. As mentioned in Section 4.4.1, the system I presented in this chapter is incorporating a crude filter based on topicality to detect similar events, but it could and should be done better. There has been recent work addressing the problem of event coreference. Wang & Zhang (2009) propose addressing event coreference using predicate-argument structure. They also confirm that filtering coreferent from non-coreferent pairs, as I suggest in section 4.4, does improve performance. They show that in the three-way textual inference task (in which systems are required to tag pairs as entailed, contradictory or unknown), systems are better off distinguishing first between coreferent and non-coreferent pairs rather than making the first distinction at the level of inference or contradiction (3 to 6% improvement in accuracy, depending on the dataset). Bejan & Harabagiu (2010) also targets event coreference, and developed a corpus of cross-document event coreference (the EventCorefBank which is available at http://faculty.washington.edu/bejan/). Lee *et al.* (2012) extend the Bejan and Harabagiu corpus, labeling both entities and events across documents. They propose to jointly model entity and event coreference, and show that when allowing information to flow between both types of coreference, event coreference helps improve entity coreference, and vice versa.

Kim & Zhai (2009) suggest an application of contradiction detection: they propose a new task crossing the domains of summarization and sentiment analysis. For product reviews, they generate a list of contrastive pairs of sentences with different sentiment polarities to help the reader digest contradictory opinions on the same product. To do so, they need to first identify pairs of sentences related to similar product features, and then among these, find the contrastive ones. Paul *et al.* (2010) further work on summarizing contrastive viewpoints in opinionated texts. They propose two types of summaries: macro- and micro-level summaries. The macro summaries consist of multiples sets of sentences, each representing a different viewpoint, whereas the micro summaries are closer to the work of Kim & Zhai (2009): they consist of pairs of sentences which stand in contrast.

Uptakes were not confined to the computational linguistics world. In their work on the interpretation of appositives, Harris & Potts (2009) refer to the contradiction patterns I identified (see section 4.3). They built a corpus of embedded appositives to assess how such appositives in naturally-occurring text are interpreted by readers: at the text level (such as in (94)) or at the embedded level (such as in (95)).

(94) A government prosecutor said Wednesday he plans to drop vandalism charges against a Malaysian teenager allegedly involved in a spate of spray painting cars with a young American, Michael Fay, who was caned recently.

(95) Israel says Arad was captured by Dirani, who may have then sold him to Iran.

When possible the annotation gives evidence for the interpretation. For example, in (94), the evidence is that "A later sentence elaborates on the details of Fay's punishment: [...] *was given four strokes of a rattan cane two weeks ago* [...]". Since the definition of contradiction I used is faithful to the goal of Harris & Potts, i.e., examining how readers interpret such appositive constructions, one of the kinds of evidence that was used are the patterns of contradiction I identified.

The taxonomy of patterns of contradiction I identified is one of the principal contributions of this chapter. Through a detailed analysis of naturally-occurring conflicting statements, I identified linguistic factors which give rise to contradiction, and proposed a classification of these. The main contribution of this chapter is a working definition of contradiction that targets the way people interpret texts. As I emphasized in chapter 1, the field of computational linguistics needs to move in the direction of capturing people's interpretation of language, if our goal is to build systems capable of providing language understanding comparable to what humans achieve. Detecting conflicting information is a task that we naturally perform. The definition seeks to capture the pragmatic meaning of contradiction. It is based on common-sense rather than on a logical notion of contradiction, and therefore encapsulates the two cases of conflicting information that typically arise in text: divergent facts and different opinions (see section 4.3). An important point is the high inter-annotator agreement found in the corpus annotation. The definition of contradiction that I proposed could face similar critiques to the definition of RTE, and could be viewed as vague. However such a high agreement indicates that my definition perfectly suits people's interpretation of contradiction: it conforms to readers' intuitions of what counts as contradictory.

Chapter 5

The pragmatic complexity of veridicality assessment

5.1 Introduction

In this chapter I will focus on a third kind of inferences people commonly draw in their everyday use of language: inferences made about the status of an event. When an event is described, we assess how it relates to the real world: to what extent do we believe that the speaker (author) intends to convey that the event did (or did not) happen? As emphasized in the introduction, inferences we make about the status of an event are part of the pragmatic meaning we retrieve from what we read or hear. Thus, natural language understanding depends heavily on assessing whether events mentioned in a text are viewed as happening or not. An unadorned declarative like *The cancer has spread* conveys firm speaker commitment, whereas qualified variants such as *There are strong indicators that the cancer has spread* or *The cancer might have spread* imbue the claim with uncertainty. Building on logical, linguistic, and computational insights about the relationship between language and reader commitment (Montague 1969; Barwise 1981; Giannakidou 1994; Giannakidou 1995; Giannakidou 1999; Giannakidou 2001; Zwarts 1995; Asher & Lascarides 2003; Karttunen & Zaenen 2005; Rubin *et al.* 2005; Rubin 2007; Saurí 2008), I will refer to the problem of whether events correspond to situations in the real world as *event* *veridicality.* The central goal of this chapter is to begin to identify the linguistic and contextual factors that shape readers' veridicality judgments.

There is a long tradition of tracing veridicality to fixed properties of lexical items (Kiparsky & Kiparsky 1970; Karttunen 1973). On this view, a lexical item L is *veridical* if the meaning of L applied to argument p entails the truth of p. For example, since both true and false things can be believed, one should not infer directly from *A believes that S* that S is true, making *believe* non-veridical. Conversely, *realize* appears to be veridical, because realizing S entails the truth of S. The prototypical anti-veridical operator is negation, since *not S* entails the falsity of S, but anti-veridicality is a characteristic of a wide range of words and constructions (e.g., *have yet to*, *fail to*, *without*). These basic veridicality judgments can be further subdivided using modal or probabilistic notions. For example, while *may* is non-veridical by the lexicalist classification, we might classify *may S* as *possible* with regard to S.[1]

Lexical theories of this sort provide a basis for characterizing readers' veridicality judgments, but they do not tell the whole story, because they neglect the sort of pragmatic enrichment that is pervasive in human communication. On the lexical view, *say* can only be classified as *non-veridical* (both true and false things can be said), and yet, if a *New York Times* article contained the sentence *United Widget said that its chairman resigned*, readers would reliably infer that United Widget's chairman resigned — the sentence is, in this context, veridical (at least to some degree) with respect to the event described by the embedded clause, with *United Widget said* functioning to mark the source of the information conveyed (Simons 2007). *Cognitive authority*, as termed in information science (Rieh 2010), plays a crucial role in how people judge the veridicality of events. Here, the provenance of the document (*The New York Times*) and the source (United Widget) combine to reliably lead a reader to infer that the author intended to convey that the event really happened. Conversely, *allege* is lexically non-veridical, and yet this only begins to address the complex interplay of world knowledge and lexical meaning that will shape people's inferences about the sentence *FBI agents alleged in court documents*

[1] Lexical notions of veridicality must be relativized to specific argument positions, with the other arguments existentially closed for the purposes of checking entailment. For example, *believe* is non-veridical on its inner (sentential) argument because '$\exists x : x$ believes p' does not entail p.

today that Zazi had admitted receiving weapons and explosives training from al Qaeda operatives in Pakistan last year. I conclude from examples like these that veridicality judgments have an important pragmatic component, and, in turn, that veridicality should be assessed using information from the entire sentence as well as from the context. Lexicalist theories have a significant role to play here, but I expect their classifications to be altered by other communicative pressures. For example, the lexical theory can tell us that, as a narrowly semantic fact, X *alleges* S is non-veridical with regard to S. However, where X is a trustworthy source for S-type information, we might fairly confidently conclude that S is true. Where X is known to spread disinformation, we might tentatively conclude that S is false. These pragmatic enrichments move us from uncertainty to some degree of certainty. Such enrichments can be central to understanding the speaker's (or author's) intended message.

Embracing pragmatic enrichment means embracing uncertainty. While speakers can feel reasonably confident about the core lexical semantics of the words of their language, there is no such firm foundation when it comes to this kind of pragmatic enrichment. The newspaper says, *United Widget said that its profits were up in the fourth quarter*, but just how trustworthy is United Widget on such matters? Speakers are likely to vary in what they intend in such cases, and hearers (readers) are thus forced to operate under uncertainty when making the requisite inferences. Lexical theories allow us to abstract away from these challenges, but a pragmatically-informed approach must embrace them.

The FactBank corpus is a leading resource for research on veridicality (Saurí & Pustejovsky 2009). Its annotations are "textual-based": they seek to capture the ways in which lexical meanings and local semantic interactions determine veridicality judgments. In order to better understand the role of pragmatic enrichment, I had a large group of linguistically naive annotators annotate a portion of the FactBank corpus, given very loose guidelines. Whereas the FactBank annotators were explicitly told to avoid bringing world knowledge to bear on the task, my annotators were encouraged to choose labels that reflected their own natural reading of the texts. Each sentence was annotated by 10 annotators, which increases my confidence in the annotation and also highlights the sort of vagueness and ambiguity that can affect veridicality. These new annotations help confirm my hypothesis that veridicality judgments are shaped by a variety of other linguistic and contextual factors beyond lexical meanings.

The nature of contextual factors is central to linguistic pragmatics, and as pointed out by Karttunen & Zaenen (2005), veridicality is fundamental to a range of natural language processing tasks, including information extraction, opinion detection, and textual entailment. Veridicality is prominent in BioNLP, where identifying negations (Chapman *et al.* 2001; Elkin *et al.* 2005; Huang & Lowe 2007; Pyysalo *et al.* 2007; Morante & Daelemans 2009) and hedges or "speculations" (Szarvas *et al.* 2008; Kim *et al.* 2009) is crucial to proper textual understanding. Recently, more attention has been devoted to veridicality within NLP, with the 2010 workshop on negation and speculation in natural language processing (Morante & Sporleder 2010) and a special issue in *Computational Linguistics* (Morante & Sporleder 2012). Veridicality was also at the heart of the 2010 CoNLL shared task (Farkas *et al.* 2010), where the goal was to distinguish uncertain events from the rest. However, the centrality of veridicality assessment to tasks like event and relation extraction is arguably still not fully appreciated. At present the vast majority of information extraction systems work at roughly the clause level and regard any relation they find as true. But relations in actual text may not be facts for all sorts of reasons, such as being embedded under an attitude verb like *doubt*, being the antecedent of a conditional, or being part of the report by an untrustworthy source. To avoid wrong extractions in these cases, it is essential for NLP systems to assess the veridicality of extracted facts.

I argue for three main claims about veridicality. First and foremost, I aim to show that pragmatically informed veridicality judgments are systematic enough to be included in computational work on textual understanding. Second, I seek to justify FactBank's seven-point categorization of veridicality over simpler alternatives (e.g., certain vs. uncertain, as in the CoNLL task). Finally, the inherent uncertainty of pragmatic inference suggests that veridicality judgments are not always categorical, and thus are better modeled as probability distributions over veridicality categories. To substantiate these claims, I analyze in detail the annotations I collected, and I report on experiments that treat veridicality assessment as a distribution-prediction task. The feature set I use includes not only lexical items like hedges, modals, and negations, but also complex structural features and approximations of world knowledge. Though the resulting classifier is limited in being able to assess veridicality in complex real world contexts, it still does quite a good job of capturing human pragmatic judgments of veridicality. I argue that the model yields insights into the complex pragmatic factors that shape readers' veridicality judgments.

5.2 Veridicality and modality

The concept of veridicality is related to the one of factuality. Saurí defines factuality as "the level of information expressing the commitment of relevant sources towards the factual nature of eventualities in text. That is, it is in charge of conveying whether eventualities are characterized as corresponding to a fact, to a possibility, or to a situation that does not hold in the world" (Saurí 2008:1). However the term 'veridicality' better describes the issue I am interested in: the use of the term 'veridicality' closely matches that of Giannakidou (1999), where it is defined so as to be (i) relativized to particular agents or perspectives, (ii) gradable, and (iii) general enough to cover not only facts but also the commitments that arise from using certain referential expressions and aspectual morphology. The more familiar term 'factuality' seems at odds with all three of these criteria, so I avoid it.

Veridicality and factuality are related to the notion of modality. Modality has been widely studied and different definitions have been proposed. Lyons (1977) defines 'epistemic modality' as "the speaker's opinion or attitude towards the proposition that the sentence expresses or the situation that the proposition describes". Palmer (1986) makes a distinction between 'propositional modality' which is "concerned with the speaker's attitude to the truth-value or factual status of the proposition" (e.g., *Kate must be at home now*) and 'event modality' which "refers to events that are not actualized, events that have not taken place but are merely potential" (e.g., *Kate must come in now*). von Fintel (2006) defines modality as "a category of linguistic meaning having to do with the expression of possibility and necessity". Portner (2009) defines it as "the linguistic phenomenon whereby grammar allows one to say things about, or on the basis of, situations which need not be real". These last two definitions characterize modality as a semantic phenomenon. A wide range of linguistic expressions have (or have been claimed to have) modal semantics: modal auxiliaries, conditionals, *because*-clause, imperfective verbs, the future tense, expressions of mood, evidentials, attitude verbs. From a formal semantics point of view, modality has been defined as quantification over possible worlds (Kratzer 1981; Kratzer 1991). However, Lassiter (2011) has argued that modals are more similar in their semantics to gradable adjectives than they are to quantifiers, and propose a new approach according to which modals map propositions to scales of probability-weighted preferences and compare them to a threshold value. He shows that probability plays a crucial role in the semantics of modality. My approach will be similar: I argue that adopting a probabilistic model will more accurately represent human pragmatic judgments of veridicality. From a computational linguistics perspective, research on modality has primarily been focused on concepts which all play a role in veridicality assessment: hedging, evidentiality (information source of a statement), detection of speculative language, detection of subjective language (i.a., Wiebe 1994; Wiebe *et al.* 2004; Wiebe & Cardie 2005; Wilson *et al.* 2005; Wilson 2008; Szarvas *et al.* 2008; Szarvas 2008). As noted by Morante & Sporleder (2012), these concepts are related to the attitude of the speaker/hearer towards the statement in terms of degree of certainty, reliability, subjectivity, sources of information and perspective.

5.3 Corpus annotation

I aim at identifying linguistic and contextual factors that play a role in veridicality assessment. As mentioned above, the FactBank corpus is the primary resource to work on veridicality. Its annotations are intended to differentiate semantic effects from pragmatic ones in the area of veridicality assessment. My overarching goal is to examine how pragmatic enrichment affects this picture. Thus, I use the FactBank sentences in my own investigation, to facilitate comparisons between the two kinds of information and to create a supplement to FactBank itself. The present section

Value	**Definition**	**Count**
CT+	According to the source, it is certainly the case that X	7,749 (57.6%)
PR+	According to the source, it is probably the case that X	363 (2.7%)
PS+	According to the source, it is possibly the case that X	226 (1.7%)
CT-	According to the source, it is certainly not the case that X	433 (3.2%)
PR-	According to the source it is probably not the case that X	56 (0.4%)
PS-	According to the source it is possibly not the case that X	14 (0.1%)
CTu	The source knows whether it is the case that X or that not X	12 (0.1%)
Uu	The source does not know what the factual status of the event is, or does not commit to it	4,607 (34.2%)
		13,460

Table 5.1: FactBank annotation scheme.

introduces FactBank in more detail and then thoroughly reviews my own annotation project and its results.

5.3.1 FactBank corpus

FactBank provides veridicality annotations on events relative to each participant involved in the discourse. It consists of 208 documents from newswire and broadcast news reports in which 9,472 events (verbs, nouns, and adjectives) were manually identified. There is no fundamental difference in the way verbs, nouns, and adjectives are annotated. Events are single words. The data comes from TimeBank 1.2 and a fragment of AQUAINT TimeML (Pustejovsky *et al.* 2006). The documents in the AQUAINT TimeML subset come from two topics: “NATO, Poland, Czech Republic, Hungary” and “the Slepian abortion murder”.

The tags annotate ⟨event, participant⟩ pairs in sentences. The participant can be anyone mentioned in the sentence as well as its author. In (96), for example, the target event *means* is assigned a value with respect to both the source *some experts* and the author of the sentence.

(96) Some experts now predict Anheuser's entry into the fray **means** near-term earnings trouble for all the industry players.

Veridicality(means, experts) = PR+

Veridicality(means, author) = Uu

The tags are summarized in table 5.1. Each tag consists of a veridicality value (certain 'CT', probable 'PR', possible 'PS', underspecified 'U') and a polarity value (positive '+', negative '-', unknown 'u'). CT+ corresponds to veridicality as described in the introduction of this chapter, Uu to non-veridicality, and CT- to anti-veridicality. The PR and PS categories add a modal or probabilistic element to the scale, to capture finer-grained intuitions.

Examples (97) and (98) illustrate the annotations for a noun and an adjective.

(97) But an all-out bidding **war** between the world's top auto giants for Britain's leading luxury-car maker seems unlikely.

Veridicality(war, author) = PR-

(98) Recently, analysts have said Sun also is **vulnerable** to competition from International Business Machines Corp., which plans to introduce a group of workstations early next year, and Next Inc.

Veridicality(vulnerable, analysts) = CT+

Veridicality(vulnerable, author) = Uu

The last column of table 5.1 reports the value counts in the corpus. The data are heavily skewed, with 62% of the events falling to the positive side and 57.6% in the CT+ category alone. The inter-annotator agreement for assigning veridicality tags was high ($\kappa = 0.81$, a conservative figure given the partial ordering in the tags). A training/test split is defined in FactBank: the documents from TimeBank 1.2. are used as the training data and the ones from the subset of the AQUAINT TimeML corpus as testbed.

As noted above, FactBank annotations are supposed to be as purely semantic as possible; the goal of the project was to "focus on identifying what are the judgments that the relevant participants make about the factuality nature of events, independently from their intentions and beliefs, and exclusively based on the linguistic expressions employed in the text to express such judgements", disregarding "external factors such as source reliability or reader bias" (Saurí 2008:5). The annotation manual contains an extensive set of discriminatory tests (Saurí 2008:230–235) that are informed by lexical theories of veridicality. The resulting annotations are "textual-based, that is, reflecting only what is expressed in the text and avoiding any judgment based on individual knowledge" (Saurí & Pustejovsky 2009:253). In addition, discourse structure is not taken into account: "we decided to constrain our annotation to information only present at the sentence level" (Saurí & Pustejovsky 2009:253).

5.3.2 Annotations from the reader's perspective

FactBank seeks to capture aspects of literal meaning, whereas I aim to capture aspects of pragmatic meaning, which brings us closer to characterizing the amount and kind of information that a reader can reliably extract from an utterance. I thus extend the FactBank annotations by bringing world knowledge into the picture. Whereas the FactBank annotators were explicitly told to avoid any reader bias, to disregard the credibility of the source, and to focus only on the linguistic terms used in the text to express veridicality, I am interested in capturing how people judge the veridicality of events when reader bias, credibility of the source, and what we know about the world is allowed to play a role.

To do this, I took a subset of the FactBank sentences annotated at the author level and recruited annotators for them using Mechanical Turk. I restricted the task to annotators located in the United States. My subset consists of 642 sentences (466 verbs, 155 nouns, 21 adjectives); I use all the PR+, PS+, PR-, PS- items from the FactBank training set plus some randomly chosen Uu, CT+ and CT- items. (I did not take any CTu sentences into account, as there are not enough of them to support experimentation.) The annotators were asked to decide whether the bold-faced event described in the sentence did (or will) happen. I used Saurí's 7-point annotation scheme (removing CTu). To ensure that the workers understood the task, I first gave them four training items — simple non-corpus examples designed to help them conceptualize the annotation categories properly. The sentences were presented in blocks of 26 items, three of which were "tests" very similar to the training items, included to ensure that the workers were careful. I discarded data from two Turkers because they did not correctly tag the three test sentences.[2]

Like Saurí, I did not take the discourse structure into account: Turkers saw only disconnected sentences and judged the event sentence by sentence. Subsequent mentions of a same event in the discourse can however lead to revise a veridicality judgment already posed for that event. For instance in (99) from (Saurí 2008:56), a reader's veridicality judgment about the *tipped off* event will probably change when reading the second sentence.

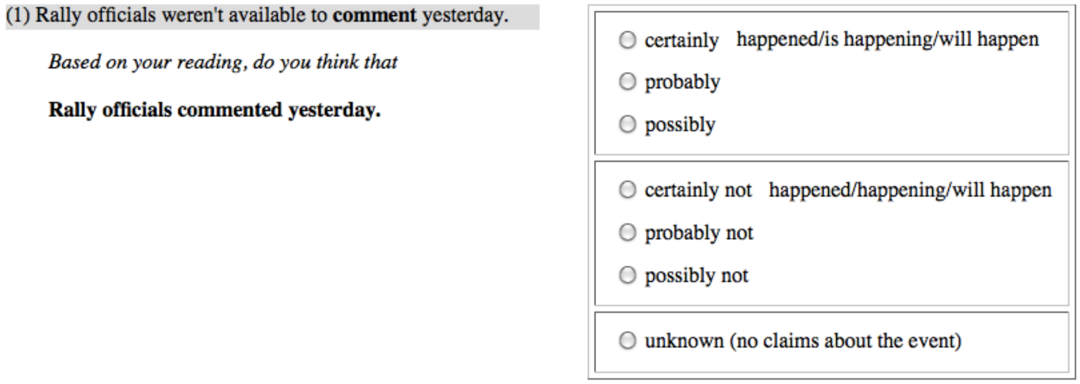
(99) Yesterday, the police denied that drug dealers were **tipped off** before the operation. However, it emerged last night that a reporter from London Weekend Television unwittingly **tipped off** residents about the raid when he phoned contacts on the estate to ask if there had been a raid — before it had actually happened.

Here, though, I concentrate on the sentence level, and leave the issue of discourse structure for future work. In other words, I capture the reader's judgment about the veridicality of an event after each sentence, independent on whether the judgment will be revised when later information is read. This is partly to facilitate comparisons with FactBank and partly because I am presently unsure how to computationally model the effects of context in this area.

Figure 5.1 shows how the items were displayed. I rephrased the event under consideration (the bold sentence in Figure 5.1), since it is not always straightforward to identify the intended rephrasing. Following Saurí, I refer to this rephrasing process as *normalization*. The normalization strips out any polarity and modality markers to

[2] The data are available at http://christopherpotts.net/ling/data/factbank/.

focus only on the core event talked about. For example, in *Police* ***gave*** *no details*, I needed to make sure that workers evaluated the positive form ("Police gave details"), rather than the negative one ("Police gave no details"). Similarly, in *Hudson's Bay Co. announced terms of a previously proposed rights issue that is expected to* ***raise*** *about 396 million Canadian dollars (US$337 million) net of expenses*, the normalization will remove the modality marker *is expected* ("the proposed rights issue will raise about 396 million Canadian dollars net of expenses"). I followed Saurí's extensive guidelines for this rephrasing process (Saurí 2008:218–222).



(1) Rally officials weren't available to **comment** yesterday.

Based on your reading, do you think that

Rally officials commented yesterday.

○ certainly happened/is happening/will happen
○ probably
○ possibly

○ certainly not happened/happening/will happen
○ probably not
○ possibly not

○ unknown (no claims about the event)

Figure 5.1: Design of the Mechanical Turk experiment.

	κ	p **value**
CT+	0.63	< 0.001
CT-	0.80	< 0.001
PR+	0.41	< 0.001
PR-	0.34	< 0.001
PS+	0.40	< 0.001
PS-	0.12	< 0.001
Uu	0.25	< 0.001
Overall	0.53	< 0.001

Table 5.2: Fleiss kappa scores with associated p-values.

I collected ten annotations for each event. 177 workers participated in the annotations. Most workers did just one batch of 23 non-test examples; the mean number of annotations per worker was 44, and they each annotated between 23 and 552 sentences. Table 5.2 reports Fleiss kappa scores (Fleiss 1971) using the full seven-category scheme. These scores are conservative since they do not take into account the fact that the scale is partially ordered, with CT+, PR+, and PS+ forming a 'positive' category, CT-, PR-, and PS- forming a 'negative' category, and Uu remaining alone. The overall Fleiss kappa for this three-category version is much higher (0.66), reflecting the fact that many of the disagreements were about degree of confidence (e.g., CT+ vs. PR+) rather than the basic veridicality judgment of 'positive', 'negative', or 'unknown'. At least 6 out of 10 workers agreed on the same tag for 500 of the 642 sentences (78%). For 53% of the examples, at least 8 Turkers agreed with each other, and total agreement is obtained for 26% of the data (165 sentences).

5.3.3 An alternative scale

One of my goals is to assess whether FactBank's seven-category scheme is the right one for the task. To this end, I also evaluated whether a five-tag version would increase agreement and perhaps provide a better match with readers' intuitions. Logically, PR- is equivalent to PS+, and PS- to PR+, so it seemed natural to try to collapse them into a two-way division between 'probable' and 'possible'. I thus ran the MTurk experiment again with the five-point scheme in table 5.3.

Category	**Original**
yes	CT+
probable	PR+/PS-
possible	PS+/PR-
no	CT-
unknown	Uu

Table 5.3: An alternative five-tag annotation scheme.

The five-point scheme led to lower agreement between Turkers. Globally, the PR- items were generally mapped to 'no', and PS- to either 'no' or 'unknown'. Some Turkers chose the expected mappings (PS- to 'probable' and PR- to 'possible'), but only very rarely. This is explicable in terms of the pragmatics of veridicality judgments. Though PR- may be logically equivalent to PS+, and PS- to PR+, there are important pragmatic differences between giving a positive judgment and giving a negative one. For example, in (100), speaker B will not infer that he can possibly get a further discount, even if 'Probably not' is consistent with 'Possibly'. Conversely, had the answer been 'Possibly', A would have remained hopeful.

(100) A: Is it possible to get further discount on the rate?
B: Probably not.

This is in line with results of experiments looking at how people interpret terms which invoke probabilities, such as *definite*, *certain*, *possible*, *probable*, *likely*. Reyna (1981) examine these terms and their negations (*indefinite*, *uncertain*, *impossible*, *improbable*, *unlikely*), as well as negating the terms with *not* (*not probable*, *not possible*, etc.). She shows that people make a distinction between possibility and probability, and that the impact of negation is important. Mosteller & Youtz (1990) asked subjects to attribute probabilities to such terms, and also show that people clearly make a distinction between the terms *certain*, *probable* and *possible* (there is no overlap in the probability values subjects assigned to these terms). In sum, there seems to be a very intuitive notion of veridicality along the partially ordered scale proposed by Saurí.

In their work on assessing the degree of event certainty to which an author commits, Rubin *et al.* (2005) used the following five-point scale: *absolute*, *high*, *moderate*, *low*, and *uncertain*. They did not obtain very high inter-annotator agreement ($\kappa = 0.41$). Saurí hypothesized that their low agreement is due to a fuzzy approach and the lack of precise guidelines. Rubin *et al.* (2005) had no clear identification of certainty markers, and no explicit test for distinguishing different degrees of certainty (unlike Saurí). However, in my experiment, the guidelines were similarly loose: Turkers were instructed only to "read 30 sentences and decide whether the events described in these sentences did (or will) happen". They were not asked to limit their attention to just the information in the sentence, and they were not given any mappings between linguistic markers and veridicality values. Nonetheless, Turkers reached good agreement levels in assessing event veridicality. I conclude from this that Saurí's scale comes closer than its competitors to capturing speaker intuitions about veridicality. This mirrors the general high inter-annotator agreement levels that have been found for the Recognizing Textual Entailment task (Manning 2006) as well as what I found for contradiction detection (see section 4.2.2), perhaps reflecting that judging inference, contradiction and veridicality in context is a natural, everyday human task.

Diab *et al.* (2009) annotated a 10,000-word corpus for what they call "committed beliefs": whether the author of the sentence indicates with linguistic means that he believes or disbelieves that the event described by the sentence is a fact. Thus, in essence, the annotations assess the degree of event certainty to which an author commits, as in Rubin's work. They employ a 3-point scale: *committed belief*, *non-committed belief*, and *not applicable*. An example of *committed belief* is *GM has laid off workers*. Affirmative sentences in the future are also considered as *committed belief* (e.g., *GM will lay off workers*). Sentences with modals and events embedded under speech verbs are annotated as *non-committed belief*. The third category, *not applicable*, consists of sentences expressing desire (*Some wish GM would lay of workers*), questions (*Many wonder if GM will lay off workers*), and requirements (*Lay off workers!*). The corpus covers different genres (newswire, email, blog, dialogue). The inter-annotator agreement was high (95%). Prabhakaran *et al.* (2010) used the corpus to automatically tag committed beliefs according to that 3-point scale. This too is an important resource, but it is difficult to compare it with my own task, for two reasons. First, the annotators sought to prevent world knowledge from influencing their annotations, which is concerned only with linguistic markers. Second, the category *non-committed belief* conflates the possible, probable, and unknown categories of my corpus (Saurí's). Though some work in the biomedical domain (i.a., Hobby *et al.* (2000)) suggests that the distinction between possible and probable is hard to make, I did not want to avoid it, since people routinely make such fine-grained modal distinctions when assessing claims. What's more, the approach I develop allows me to quantify the degree to which such judgments are in fact variable and uncertain.

	κ	p **value**
CT+	0.37	< 0.001
PR+	0.79	< 0.001
PS+	0.86	< 0.001
CT-	0.91	< 0.001
PR-	0.77	< 0.001
PS-	-0.001	$= 0.982$
Uu	0.06	$= 0.203$
Overall	0.60	< 0.001

Table 5.4: Inter-annotator agreement comparing FactBank annotations with MTurk annotations. The data are limited to the 500 examples in which at least 6 of the 10 Turkers agreed on the label, which is then taken to be the true MTurk label. The very poor value for PS- derives from the fact, in this subset, that label was chosen only once in FactBank and not at all by our annotators.

5.4 Lessons from the new annotations

This section presents two kinds of high-level analysis of my annotations. I first compare them with FactBank' annotations for veridicality according to the author, identifying places where the annotations point to sharp divergences between literal meaning and pragmatic meaning. I then study the full distribution of annotations I received (10 per sentence), using them to highlight the uncertainty of veridicality judgments. Both of these discussions deeply inform the modeling work presented in section 5.5.

5.4.1 The impact of pragmatic enrichment

Although the MTurk annotations largely agree with those of FactBank, there are systematic differences between the two that are indicative of the ways in which pragmatic enrichment plays a role in assessing veridicality. The goal of this section is to uncover those differences. To sharpen the picture, I limit attention to the sentences for which there is a majority-vote category, i.e., at least six out of ten Turkers annotated the event with the same veridicality value. This threshold was met for 500 of the 642 examples.

Table 5.4 uses kappa scores to measure the agreement between FactBank and our annotations on this 500-sentence subset of the data. I treat FactBank as one annotator and the collective Turkers as a second annotator, with the majority label the correct one for that annotator. What we see is modest to very high agreement for all the categories except Uu. The agreement level is also relatively low for CT+. The corresponding confusion matrix in table 5.5 helps explicate these numbers. The Uu category is used much more often in FactBank than by Turkers, and the dominant alternative choice for the Turkers was CT+. Thus, the low score for Uu also effectively drops the score for CT+. The question is why this contrast exists. Why do Turkers choose CT+ where FactBank says Uu?

Fact-				MTurk				
Bank	CT+	PR+	PS+	CT-	PR-	PS-	Uu	Total
CT+	54	2	0	0	0	0	0	56
PR+	4	63	2	0	0	0	0	69
PS+	1	1	55	0	0	0	2	59
CT-	5	0	0	146	0	0	2	153
PR-	0	0	0	0	5	0	1	6
PS-	0	0	0	0	0	0	1	1
Uu	94	18	9	12	2	0	21	156
Total	158	84	66	158	7	0	27	500

Table 5.5: Confusion matrix comparing the FactBank annotations (rows) with our annotations (columns).

The divergence can be traced to the way in which lexicalist theories handle events embedded under attitude predicates like *say*, *report*, and *indicate*: any such embedded event is tagged Uu in FactBank. In the MTurk annotations, readers are not viewing the veridicality of reported events as unknown. Instead they are sensitive to a wide range of syntactic and contextual features, including markers in the embedded clause, expectations about the subject as a source for the information conveyed by the embedded clause, and lexical competition between the author's choice of attitude predicate and its alternatives. For example, even though the events in (101) are all embedded under an attitude predicate (*say*), the events in (101a) and (101b) are assessed as certain (CT+), whereas the words *highly confident* in (101c) trigger PR+, and *may* in (101d) leads to PS+.

(101) a. Magna International Inc.'s chief financial officer, James McAlpine, **resigned** and its chairman, Frank Stronach, is stepping in to help turn the automotive-parts manufacturer around, the company said.

Normalization: James McAlpine resigned

Annotations: CT+: 10

b. In the air, U.S. Air Force fliers say they have **engaged** in "a little cat and mouse" with Iraqi warplanes.

Normalization: U.S. Air Force fliers have engaged in "a little cat and mouse" with Iraqi warplanes

Annotations: CT+: 9, PS+: 1

c. Merieux officials said last week that they are "highly confident" the offer will be **approved**.

Normalization: the offer will be approved

Annotations: PR+: 10

d. U.S. commanders said 5,500 Iraqi prisoners were taken in the first hours of the ground war, though some military officials later said the total may have **climbed** above 8,000.

Normalization: the total Iraqi prisoners climbed above 8,000

Annotations: PS+: 7, PR+: 3

In (101a), *the company said* is a parenthetical modifying the main clause (Ross 1973). Asher (2000), Rooryck (2001), and Simons (2007) argue that such constructions often mark the evidential source for the main-clause information, rather than embedding it semantically. In terms of my annotations, this predicts that CT+ or CT- judgments will be common for such constructions, because they become more like two-part meanings: the main-clause and the evidential commentary.

To test this hypothesis, I took from the Turkers annotations the subset of sentences tagged Uu in FactBank where the event is directly embedded under an attitude verb or introduced by a parenthetical. I removed examples where a modal auxiliary modified the event, because those are a prominent source of non-CT annotations independently of attitude predication. This yielded a total of 78 sentences. Of these, 33 are parenthetical, and 31 (94%) of those are tagged CT+ or CT-. Of the remaining 45 non-parenthetical examples, 42 (93%) of those are tagged CT+ or CT-. Thus, both parenthetical and non-parenthetical verbs are about equally likely to lead to a CT tag.[3]

This finding is consistent with the evidential analysis of such parentheticals, but it suggests that standard embedding can function pragmatically as an evidential as well. This result is expected under the analysis of Simons (2007) (see also Frazier & Clifton (2005); Clifton & Frazier (2010)). It is also anticipated by Karttunen (1973), who focuses on the question of whether attitude verbs are *plugs* for presuppositions, that is, whether presuppositions introduced in their complements are interpreted as semantically embedded. He reviews evidence suggesting that these verbs can be veridical with respect to such content, but he tentatively concludes that these are purely pragmatic effects, writing, "we do not seem to have any alternative except to classify all propositional attitude verbs as plugs, although I am still not convinced that this is the right approach" (Karttunen 1973:190). The evidence presented here leads me to agree with Karttunen about this basic lexicalist classification, with the major caveat that the pragmatic meanings involved are considerably more complex. (For additional discussion of this point, see (Zaenen *et al.* 2005; Manning 2006).)

There are also similarities between the FactBank annotations and the Turkers

[3] I do not regard this as evidence that there is no difference between parenthetical and non-parenthetical uses when it comes to veridicality, but rather only that the categorical examples do not reveal one. Indeed, if we consider the full distribution of annotations, then a linear model with Parenthetical and Verb predicting the number of CT tags reveals Parenthetical to be a modest but significant positive predictor (coefficient estimate = 1.36, $p = 0.028$).

annotations in the case of Uu. As in FactBank, antecedents of conditionals (102), generic sentences (103), and clear cases of uncertainty with respect to the future (104) were tagged Uu by a majority of Turkers.

(102) a. If the heavy outflows **continue**, fund managers will face increasing pressure to sell off some of their junk to pay departing investors in the weeks ahead.

Normalization: the heavy outflows will continue

Annotations: Uu: 7, PS+: 2, CT+: 1

b. A unit of DPC Acquisition Partners said it would seek to liquidate the computer-printer maker "as soon as possible," even if a merger isn't **consummated**.

Normalization: a merger will be consummated

Annotations: Uu: 8, PS+: 2

(103) When prices are tumbling, they must be **willing** to buy shares from sellers when no one else will.

Normalization: they are willing to buy shares

Annotations: Uu: 7, PR+: 2, PS+: 1

(104) a. The program also calls for **coordination** of economic reforms and joint improvement of social programs in the two countries.

Normalization: there will be coordination of economic reforms and joint improvement of social programs in the two countries

Annotations: Uu: 7, PR+: 2, PS+: 1

b. But weak car sales raise questions about future **demand** from the auto sector.

Normalization: there will be demand from the auto sector

Annotations: Uu: 6, PS+: 2, CT+: 1, PR-: 1

Another difference between FactBank and the Turkers is the more nuanced categories for PS and PR events. In FactBank, markers of possibility or probability, such as *could* or *likely*, uniquely determine the corresponding tag (Saurí 2008:233). In contrast, the Turkers allow the bias created by these lexical items to be swayed by other factors. For example, the auxiliary *could* can trigger a possible or an unknown event (105). In FactBank, all these sentences are marked PS+.

(105) a. They aren't being allowed to leave and could **become** hostages.

Normalization: they will become hostages

Annotations: PS+: 10

b. Iraq could start **hostilities** with Israel either through a direct attack or by attacking Jordan.

Normalization: there will be hostilities

Annotations: Uu: 6, PS+: 3, PR+: 1

Similarly, *expected* and *appeared* are often markers of PR events. However, whereas it is uniquely so in FactBank, our annotations show a lot of shifting to PS. Examples (106) and (107) highlight the contrast: it seems likely that the annotators simply have different overall expectations about the forecasting described in each example, a high-level pragmatic influence that does not attach to any particular lexical item.

(106) a. Big personal computer makers are developing 486-based machines, which are expected to **reach** the market early next year.

Normalization: 486-based machines will reach the market early next year

Annotations: PR+: 10

b. Beneath the tepid news-release jargon lies a powerful threat from the brewing giant, which last year accounted for about 41% of all U.S. beer sales and is expected to see that **grow** to 42.5% in the current year.

Normalization: there will be growth to 42.5% in the current year

Annotations: PS+: 6, PR+: 3, CT+: 1

(107) a. Despite the lack of any obvious successors, the Iraqi leader's internal power base appeared to be **narrowing** even before the war began.

Normalization: the Iraqi leader's internal power base was narrowing even before the war began

Annotations: PR+: 7, CT+: 1, PS+: 1, PS-: 1

b. Saddam appeared to **accept** a border demarcation treaty he had rejected in peace talks following the August 1988 cease-fire of the eight-year war with Iran.

Normalization: Saddam accepted a border demarcation treaty

Annotations: PS+: 6, PR+: 2, CT+: 2

Another difference is that nouns appearing in a negative context were tagged as CT+ by the Turkers but as CT- or PR- in FactBank:

(108) However, its equity in the net income of National Steel declined to $6.3 million from $10.9 million as a result of softer demand and lost **orders** following prolonged labor talks and a threatened strike.

Normalization: there were orders

Annotations: CT+: 6, PR+: 1, PR-: 1, PS+: 1, Uu: 1

This seems to trace to uncertainty about what the annotation should be when the event involves a change of state (from orders existing to not existing). Saurí & Pustejovsky (2009:260) note that noun events were a frequent source of disagreement between the two annotators since the annotation guidelines did not address at all how to deal with them.

5.4.2 The uncertainty of pragmatic enrichment

For the purposes of comparing the Turkers annotations with those of FactBank, it is useful to single out the Turkers' majority-choice category, as I did above. However, I have 10 annotations for each event, which invites exploration of the full distribution of annotations, to see if the areas of stability and variation can teach us something about the nature of speakers' veridicality judgments. In this section, I undertake such an exploration, arguing that the patterns reveal veridicality judgments to be importantly probabilistic, as one would expect from a truly pragmatic phenomenon.

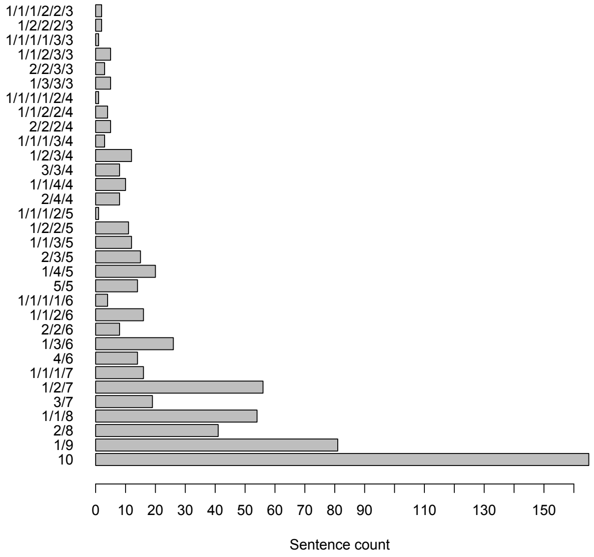


Figure 5.2: Distributions by type.

Figure 5.2 provides a high-level summary of the reaction distributions that our sentences received. The labels on the y-axis characterize types of distribution. For example, '5/5' groups the sentences for which the annotators were evenly split between two categories (e.g., a sentence for which 5 Turkers assigned PR+ and 5 assigned PS+, or a sentence for which 5 Turkers chose PR+ and 5 chose Uu). The largest grouping, '10', pools the examples on which all the annotators were in agreement.

We can safely assume that some of the variation seen in figure 5.2 is due to the noisiness of the crowd-sourced annotation process. Some annotators might have been inattentive or confused, or simply lacked the expertise to make these judgments (Snow *et al.* 2008). For example, the well-represented '1/9' and '1/1/8' groups probably represent examples for which veridicality assessment is straightforward but one or two of the annotators did not do a good job. If all the distributions were this skewed, we might feel secure in treating veridicality as categorical. However, there are many examples for which it seems implausible to say that the variation is due to noise. For example, '5/5' groups include sentences like (109) and (110), for which the judgments depend heavily on one's prior assumptions about the entities and concepts involved.

(109) In a statement, the White House said it would do "whatever is necessary" to ensure **compliance** with the sanctions.

> Normalization: there will be compliance with the sanctions
>
> Annotations: Uu: 5, PR+: 5

(110) Diplomacy appears to be making headway in **resolving** the United Nations' standoff with Iraq.

> Normalization: diplomacy is resolving the United Nations' standoff with Iraq
>
> Annotations: PR+: 5, PS+: 5

The '4/6', '1/4/5', and '4/4' groups contain many similarly difficult cases. Combining all of the rows where two categories received at least 3 votes, we get 162 examples, which is 25% of the total data set. Thus, a non-negligible subset of our sentences seem to involve examples where readers' responses are divided, suggesting that there is no unique correct label for them.

Finally, it seems likely that the long-tail of very high-entropy distributions at the top of the graph in figure 5.2 is owed in large part to the fact that veridicality judgments are often not reachable with confidence, because the utterance is inherently underspecified or because additional contextual information is needed in order to be sure. This too suggests that it would be foolhardy to assign a unique veridicality label to every example. Of course, situating the sentences in context would reduce some of this uncertainty, but no amount of background information could eliminate it entirely.

Looking more closely at these distributions, I find additional evidence for the idea that veridicality is graded and variable. One of the most striking patterns concerns the question of whether the annotators enriched an example at all, in the following sense. Consider an event that is semantically non-veridical. This could be simply because it is embedded under a non-factive attitude predicate (*say*, *allege*), or an evidential marker (*according to sources*, *it seems*). The semantic strategy for such cases is to pick Uu. However, depending on the amount and nature of the contextual information brought to bear on the assessment, one might enrich this into one of the positive or negative categories. A cautious positive enrichment would be PS+, for example.

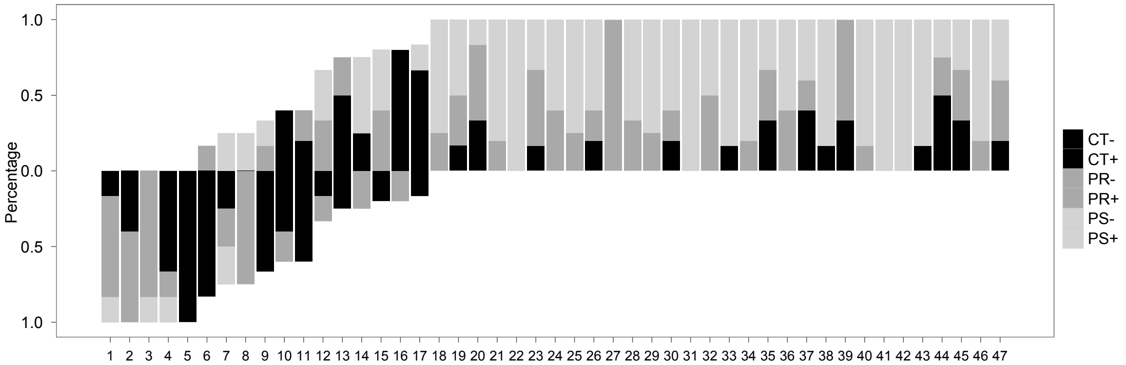


Figure 5.3: The subset of '4/6' and '5/5' distributions in which one of the dominant categories was Uu. The bars represent the distribution of non-Uu tags for each of these sentences. The top portion depicts the positive tags, and the bottom portion depicts the negative tags.

In light of this, it seems promising to look at the subset of '4/6' and '5/5' examples in which one of the chosen categories is Uu, to see what the other choices are like. On the enrichment hypothesis, the other choices should be uniformly positive or negative (up to some noise). Figure 5.3 summarizes the sentences in our corpus that result in this kind of split. The y-axis represents the percentage of non-Uu tags, with the positive values (CT+, PR+, PS+) extending upwards and the negative ones extending downwards. For sentences 1–5 and 18–47 (74% of the total), all of the non-Uu tags were uniform in their basic polarity. What's more, the distributions within the positive and negative portions are highly systematic. In the positive realm, the dominant choice is PS+, the most tentative positive enrichment, followed by PR+, and then CT+. (In the negative realm, CT- is the most represented, but I am unsure whether this supports any definitive conclusions, given the small number of examples.) My generalization about these patterns is that enrichment from a semantic Uu baseline is systematic and common, though with interesting variation both in whether it occurs and, if it does, how much.

The full distributions are also informative when it comes to understanding the range of effects that specific lexical items can have on veridicality assessment. To illustrate, I focus on the modal auxiliary verbs *can*, *could*, *may*, *might*, *must*, *will*, *would*.[4] In keeping with lexicalist theories, when they are clausemate to an event-denoting verb or event description, that event is often tagged with one of the PR and PS tags. However, the relationship is a loose one; the modal seems to steer people into these weaker categories but does not determine their final judgment. I illustrate in (111) with examples involving *may* in positive contexts.

(111) a. Last Friday's announcement was the first official word that the project was in trouble and that the company's plans for a surge in market share may have been overly **optimistic**.

Normalization: the company's plans have been overly optimistic

Annotations: PS+: 5, PR+: 5

b. In a letter, prosecutors told Mr. Antar's lawyers that because of the recent Supreme Court rulings, they could expect that any fees collected from Mr. Antar may be **seized**.

Normalization: fees collected from Mr. Antar will be seized

Annotations: PS+: 4, PR+: 6

[4] Other modals, such as *should*, *ought*, and *have to*, are either not well-represented in our data or simply absent.

c. The prospectus didn't include many details about the studio and theme park, although conceptual drawings, released this month, show that it may **feature** several "themed" areas similar to those found at parks built by Walt Disney Co.

Normalization: the parks features several "themed" areas similar to those found at parks built by Walt Disney Co.

Annotations: PS+: 4, PR+: 4, CT+:1, Uu:1

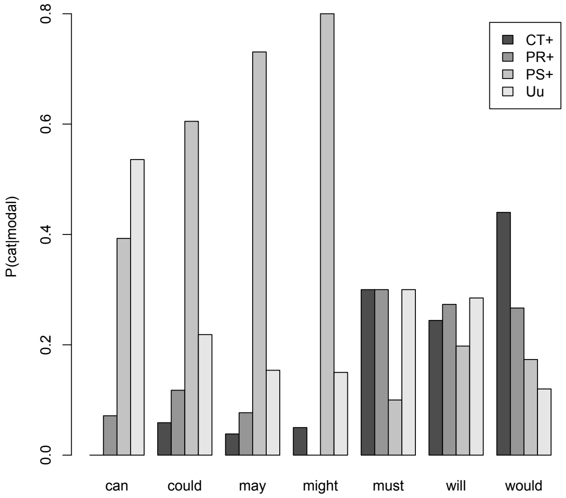


Figure 5.4: The contribution of modal auxiliaries to veridicality judgments.

Figure 5.4 summarizes the data for the full set of modals. Here, I restrict attention to event descriptions that are clausemate to a modal, effectively taking each modal to be annotated with the distribution of annotations for its clausemate event. I also look only at the positive tags, since the negative ones were too infrequent to provide reliable estimates.

Two types of modals have been recognized in the literature, *weak* and *strong* modals (Wierzbicka 1987; Sæbo 2001; von Fintel & Iatridou 2008; Finlay 2009). Each type has different distribution profiles. As expected, the weak possibility modals *can*, *could*, *may*, and *might* correlate strongly with PS. However, the other categories are also well-represented for these modals, indicating that the contribution of these markers is heavily influenced by other factors. The strong (or necessity) modals *must*, *will*, and *would* are much more evenly distributed across the categories.

The mixed picture for modal auxiliaries seems to be typical of modal markers more generally. There is not enough data to present a quantitative picture for items like *potentially*, *apparently*, and *partly*, but the following sentences suggest that they are every bit as nuanced in their contributions to veridicality.

(112) a. Anheuser-Busch Cos. said it plans to aggressively discount its major beer brands, setting the stage for a *potentially* bruising price **war** as the maturing industry's growth continues to slow.

Normalization: there will be a bruising price war

Annotations: PS+: 5, PR+: 5

b. The portfolio unit of the French bank group Credit Lyonnais told stock market regulators that it bought 43,000 shares of Cie. de Navigation Mixte, *apparently* to help **fend** off an unwelcome takeover bid for the company.

Normalization: the 43,000 shares of Cie. de Navigation Mixte will fend off an unwelcome takeover bid for the company

Annotations: PS+: 4, PR+: 4, CT:1, Uu:1

c. Nonetheless, concern about the chip may have been responsible for a decline of 87.5 cents in Intel's stock to $32 a share yesterday in over-the-counter trading, on volume of 3,609,800 shares, and *partly* **responsible** for a drop in Compaq's stock in New York Stock Exchange composite trading on Wednesday.

Normalization: concern about the chip is responsible for a drop in Compaq's stock

Annotations: PS+: 4, PR+: 4, CT+:1, PR-:1

The discussion in this section suggests that work on veridicality should embrace variation and uncertainty as being part of the characterization of veridicality, rather than trying to approximate the problem as one of basic categorization. I now turn to experiments with a system for veridicality assessment that acknowledges the multivalued nature of veridicality.

5.5 A system for veridicality assessment

In this section, I describe a maximum entropy classifier (Berger *et al.* 1996) built to automatically assign veridicality. For such classification tasks, the dominant tradition within computational linguistics has been to adjudicate differing human judgments and to assign a single class for each item in the training data. However in section 5.4.2, I reviewed the evidence in our annotations that veridicality is not necessarily categorical, in virtue of the uncertainty involved in making pragmatic judgments of this sort. In order to align with my theoretical conception of the problem as probabilistic, I treat each annotator judgment as a training item. Thus, each sentence appears 10 times in the training data.

A maximum entropy model computes the probability of each class c given the data d as follows:

$$p(c|d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

where the features f_i are indicator functions of a property Φ of the data d and a particular class c : $f_i(c, d) \equiv \Phi(d) \wedge c = c_k$. The weights λ_i of the features are the parameters of the model chosen to maximize the conditional likelihood of the training data according to the model. The maximum entropy model thus gives us a distribution over the veridicality classes, which will be the output. To assess how good the output of the model is, I will give the log-likelihood of some data according to the model. For comparison, I will also give the log-likelihood for the exact distribution from the Turkers (which thus gives an upper-bound) as well as a log-likelihood for a baseline model which uses only the overall distribution of classes in the training data (which thus gives a lower-bound).

A maximum entropy classifier is an instance of a generalized linear model with a logit link function. It is almost exactly equivalent to the standard multi-class (also called polytomous or multinomial) logistic regression model from statistics, and readers more familiar with this presentation can think of it as such. In all my experiments, I use the Stanford Classifier (Manning & Klein 2003) with a Gaussian prior (also known as L_2 regularization) set to $\mathcal{N}(0, 1)$.[5]

The features were selected through 10-fold cross-validation on the training set.

Predicate classes Saurí (2008) defines classes of predicates (nouns and verbs) that project the same veridicality value onto the events they introduce. The classes also define the grammatical relations that need to hold between the predicate and the event it introduces, since grammatical contexts matter for veridicality. Different veridicality values will indeed be assigned to *X* in *He doesn't know that X* and in *He doesn't know if X*. The classes have names like ANNOUNCE, CONFIRM, CONJECTURE, and SAY. Like Saurí, I used dependency graphs produced by the Stanford parser (Klein & Manning 2003; de Marneffe *et al.* 2006) to follow the path from the target event to the root of the sentence. If a predicate in the path was contained in one of the classes and the grammatical relation matched, I added both the lemma of the predicate as a feature and a feature marking the predicate class.

World knowledge For each verb found in the path and contained in the predicate classes, I also added the lemma of its subject, and whether or not the verb was negated. The rationale for including the subject is that, as we saw in section 5.4, readers' interpretations differ for sentences such as *The FBI said it **received** . . .* and

[5] The maximum entropy formulation differs from the standard multi-class logistic regression model by having a parameter value for each class giving logit terms for how a feature's value affects the outcome probability relative to a zero feature, whereas in the standard multi-class logistic regression model, there are no parameters for one distinguished reference class, and the parameters for other classes say how the value of a feature affects the outcome probability differentially from the reference class. Without regularization, the maximum entropy formulation is overparameterized, and the parameters are unidentifiable, but in a regularized setting, this is no longer a problem and the maximum entropy formulation then has the advantage that all classes are treated symmetrically, with a simpler symmetric form of model regularization.

Bush said he ***received*** *. . .* , presumably because of world knowledge they bring to bear on the judgment. To approximate such world knowledge, I also obtained subject–verb bigram and subject counts from the New York Times portion of GigaWord and then included log(subject–verb-counts/subject-counts) as a feature. The intuition here is that some embedded clauses carry the main point of the sentence (Frazier & Clifton 2005; Clifton & Frazier 2010; Simons 2007), with the overall frequency of the elements introducing the embedded clause contributing to readers' veridicality assessments.

General features I used the lemma of the event, the lemma of the root of the sentence, the incoming grammatical relation to the event, and a general class feature.

Modality features I used Saurí's list of modal words as features. I distinguished between modality markers found as direct governors or children of the event under consideration (e.g., *likely* in *those vital signs are not likely to* ***change***), and modal words found elsewhere in the context of the sentence (e.g., *could* in *Iraq could start hostilities with Israel either through a direct* ***attack*** *or by attacking Jordan*). Figure 5.4 provides some indication of how these will relate to the annotations.

Negation A negation feature captures the presence of linguistic markers of negative contexts. Events are considered negated if they have a negation dependency in the graph or an explicit linguistic marker of negation as dependent (e.g., simple negation (*not*), downward-monotone quantifiers (*no*, *any*), or restricting prepositions (*without*)). Events are also considered negated if embedded in a negative context (e.g., *fail*, *cancel*).

Conditional Antecedents of conditionals and words clearly marking uncertainty are reliable indicators of the Uu category. I therefore checked for events in an *if*-clause or embedded under markers such as *call for*.

Quotation Another reliable indicator of the Uu category is quotation. I generated a quotation feature if the sentence opened and ended with quotation marks, or if the

	Train	**Test**
lower-bound	−10813.97	−1987.86
classifier	−8021.85	−1324.41
upper-bound	−3776.30	−590.75

Table 5.6: Log likelihood values for the training and test data.

root subject was *we*.

In summary, the feature set allows combinations of evidence from various sources to determine veridicality. To be sure, lexical features are important, but they must be allowed to interact with pragmatic ones. In addition, the model does not presume that individual lexical items will contribute in only one way to veridicality judgments. Rather, their contributions are affected by the rest of the feature set.

5.6 Evaluation and results

As test set, I used 130 sentences from the test items in FactBank. I took all the sentences with events annotated PR+ and PS+ at the author level (there are very few), and I randomly chose sentences for the other values (CT+, CT- and Uu since the FactBank test set does not contain any PR- and PS- items). Three colleagues of mine provided the normalizations of the sentences, and the data were then annotated using Mechanical Turk, as described in section 5.3. For 112 of the 130 sentences, at least six Turkers agreed on the same value.

Table 5.6 gives log-likelihood values of the classifier for the training and test sets, along with the upper and lower bounds. The upper bound is the log-likelihood of the model that uses the exact distribution from the Turkers. The lower bound is the log-likelihood of a model that uses only the overall rate of each class in our annotations for the training data.

KL divergence provides a related way to assess the effectiveness of the classifier. The KL divergence between two distributions is an asymmetric measure of the difference between them. I use example (101d) to illustrate, repeated here in (113).

(113) U.S. commanders said 5,500 Iraqi prisoners were taken in the first hours of the ground war, though some military officials later said the total may have **climbed** above 8,000.

> Normalization: the total Iraqi prisoners climbed above 8,000
>
> Annotations: PS+: 7, PR+: 3

For that sentence, the classifier assigns a probability of 0.64 to PS+ and 0.28 to PR+, with very low probabilities for the remaining categories. It thus closely models the distribution from the Turkers (PS+: 7/10, PR+: 3/10). The KL-divergence is correspondingly low: 0.13. The KL-divergence for a classifier that assigned 0.94 probability to the most frequent category (i.e., CT+) and 0.01 to the remaining categories would be much higher: 5.76.

The mean KL divergence of the model is 0.95 (SD 1.13) for the training data and 0.81 (SD 0.91) for the test data. The mean KL divergence for the baseline model is 1.58 (SD 0.57) for the training data and 1.55 (SD 0.47) for the test data. To assess whether the classifier is a statistically significant improvement over the baseline, I use a paired two-sided t-test over the KL divergence values for the two models. The t-test requires that both vectors of values in the comparison have normal distributions. This is not true of the raw KL values, which have approximately gamma distributions, but it is basically true of the log of the KL values: for the model's KL divergences, the normality assumption is very good, whereas for the baseline model there is some positive skew. Nonetheless, the t-test arguably provides a fair way to contextualize and compare the KL values of the two models. By this test, the model improves significantly over the lower bound (two-sided $t = -11.1983$, $df = 129$, p-value $< 2.2e{-16}$).

I can also compute precision and recall for the subsets of the data where there is a majority vote, i.e., where six out of ten annotators agreed on the same label. This

	Train					**Test**				
	#	**P**	**R**	**F1**	**Base F1**	**#**	**P**	**R**	**F1**	**Base F1**
CT+	158	74.3	84.2	78.9	32.6	61	86.9	86.9	86.9	31.8
CT-	158	89.4	91.1	90.2	34.1	31	96.6	90.3	93.3	29.4
PR+	84	74.4	69.1	71.6	19.8	7	50.0	57.1	53.3	6.9
PS+	66	75.4	69.7	72.4	16.7	7	62.5	71.4	66.7	0.0
Uu	27	57.1	44.4	50.0	10.7	6	50.0	50.0	50.0	0.0
Macro-avg		74.1	71.7	72.6	22.8		69.2	71.1	70.0	13.6
Micro-avg		78.6	78.6	78.6	27.0		83.0	83.0	83.0	22.3

Table 5.7: Precision, recall and F1 on the subsets of the training data (10-fold cross-validation) and test data where there is majority vote, as well as F1 for the baseline.

allows me to give results per veridicality tag. I take as the true veridicality value the one on which the annotators agreed. The value assigned by the classifier is the one with the highest probability. Table 5.7 reports precision, recall, and F1 scores on the training and test sets, along with the number of instances in each category. None of the items in our test data were tagged with PR- or PS- and these categories were very infrequent in the training data, so we left them out. The table also gives baseline results: I used a weighted random guesser, as for the lower-bound given in Table 5.6. To test whether the system's results are significantly better than the baseline, I used McNemar's test, which assesses whether the proportions of right and wrong predictions for the two systems are significantly different. I find that the differences are significant for both the training and test sets (p-value < 0.001).

The classifier weights give insights about the interpretation of lexical markers. Some markers behave as linguistic theories predict. For example, *believe* is often a marker of probability whereas *could* and *may* are more likely to indicate possibility. But as seen in (105) and (111), world knowledge and other linguistic factors shape the veridicality of these items. The greatest departure from theoretical predictions occurs with the SAY category, which is logically non-veridical but correlates highly with certainty (CT+) in our corpus.[6] Conversely, the class KNOW, which includes *know*, *acknowledge*, *learn*, is a marker of possibility rather than certainty. The model thus shows that to account for how readers interpret sentences, the space of veridicality should be cut up differently than the lexicalist theories propose.

5.7 Error analysis

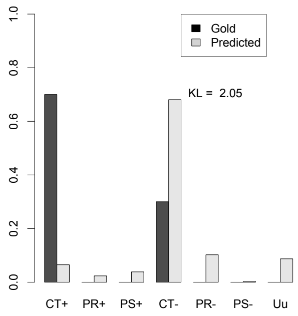
I focus on two kinds of errors. First, where there is a majority label — a label six or more of the annotators agreed on — in the annotations, I can compare that label with the one assigned the highest probability according to our model. Second, I can study cases where the the annotation distribution diverges considerably from the model's distribution (i.e., cases with a very high KL-divergence).

For the majority-label cases, errors of polarity are extremely rare; the classifier wrongly assesses the polarity of only four events, shown in (114). Most of the errors are thus in the degree of confidence (e.g., CT+ vs. PR+). The graphs next to the examples compare the annotation from the Turkers (the black bars) with the distribution proposed by the classifier (the gray bars). The KL divergence value is included to help convey how such values relate to these distributions.

(114) a. Addressing a NATO flag-lowering ceremony at the Dutch embassy, Orban said the occasion indicated the end of the embassy's **mission** of liaison between Hungary and NATO.

Normalization: there is an embassy's mission of liaison between Hungary and NATO

Annotations: CT+:7, CT-: 3

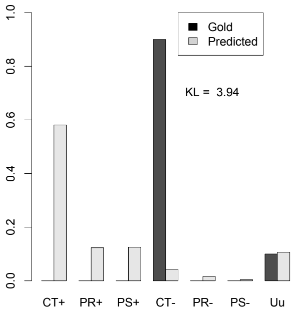


[6] This might be due to the nature of the corpus, namely newswire, where in the vast majority of the cases, reports are considered true. The situation could be totally different in another genre (such as blogs for instance).

b. But never before has NATO **reached** out to its former Eastern-bloc enemies.

Normalization: NATO has reached out to its former Eastern-bloc enemies in the past

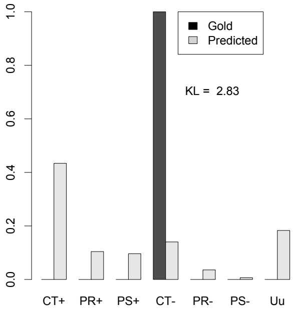
Annotations: CT-: 9, Uu: 1



c. Horsley **was** not a defendant in the suit, in which the Portland, Ore., jury ruled that such sites constitute threats to abortion providers.

Normalization: Horsley was a defendant in the suit

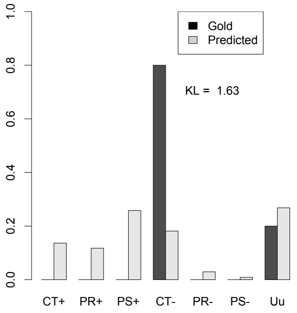
Annotations: CT-: 10



d. A total of $650,000, meanwhile, is being offered for information leading to the **arrest** of Kopp, who is charged with gunning down Dr. Barnett Slepian last fall in his home in Buffalo.

Normalization: Kopp has been arrested

Annotations: CT-: 8, Uu: 2



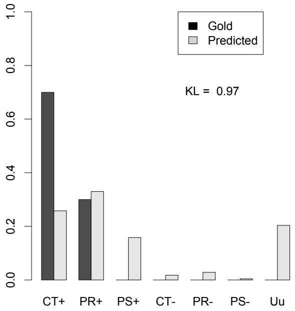
When the system missed CT- events, it failed to find an explicit negative marker, as in (114b), where, due to a parse error, *never* is treated as a dependent of the verb *have* and not of the *reaching out* event. Similarly, the system could not capture instances in which the negation was merely implicit, as in (114d), where the non-veridicality of the arresting event requires deeper interpretation than our feature-set can manage.

In (115), I give examples of CT+ events that are incorrectly tagged PR+, PS+, or Uu by the system because of the presence of a weak modal auxiliary or a verb that lowers certainty, such as *believe*. As we saw in section 5.4.2, these markers correlates strongly with the PS categories.

(115) a. The NATO summit, she said, would produce an **initiative** that "responds to the grave threat posed by weapons of mass destruction and their means of delivery."

Normalization: there will be an initiative

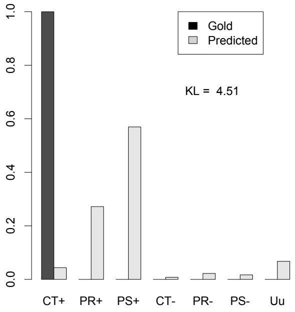
Annotations: CT+: 7, PR+: 3



b. Kopp, meanwhile, may have approached the border with Mexico, but it is **unknown** whether he crossed into that country, said Freeh.

Normalization: it is unknown whether Kopp crossed into Mexico

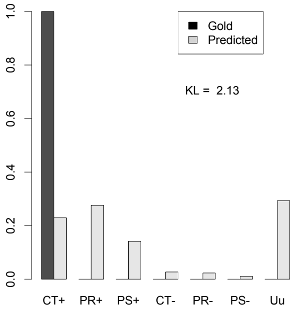
Annotations: CT+: 10



c. They believe Kopp was driven to Mexico by a female friend after the **shooting**, and have a trail of her credit card receipts leading to Mexico, the federal officials have said.

Normalization: there was a shooting

Annotations: CT+: 10

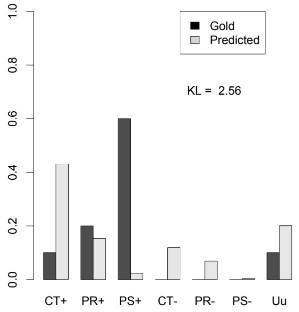


In the case of PR+ and PS+ events, all the erroneous values assigned by the system are CT+. Some explicit modality markers were not seen in the training data, such as *potential* in (116a), and thus the classifier assigned them no weight. In other cases, such as (116b), the system did not capture the modality implicit in the conditional.

(116) a. Albright also used her speech to articulate a forward-looking vision for NATO, and to defend NATO's potential **involvement** in Kosovo.

Normalization: NATO will be involved in Kosovo

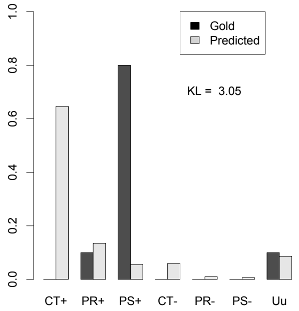
Annotations: PS+: 6, PR+: 2, CT+: 1, Uu: 1



b. "And we must be resolute in spelling out the consequences of intransigence," she added, referring to the threat of NATO air **strikes** against Milosevic if he does not agree to the deployment.

Normalization: there will be NATO air strikes

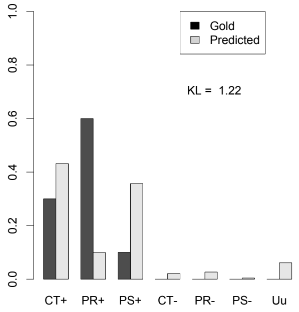
Annotations: PS+: 8, PR+: 1, Uu: 1



c. But the decision by District Attorney Frank C. Clark to begin presenting evidence to a state grand jury suggests that he has **amassed** enough material to support a criminal indictment for homicide.

Normalization: District Attorney Frank C. Clark has amassed material to support a criminal indictment for homicide

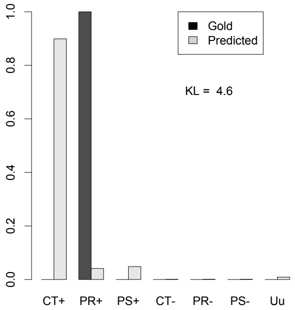
Annotations: PR+: 6, CT+: 3, PS+: 1



d. The first round of DNA tests on the hair at the FBI Laboratory here established a high probability it **came** from the same person as a hair found in a New Jersey home where James C. Kopp, a 44-year-old anti-abortion protester, lived last year, the official said.

Normalization: the hair came from the same person

Annotations: PR+: 10

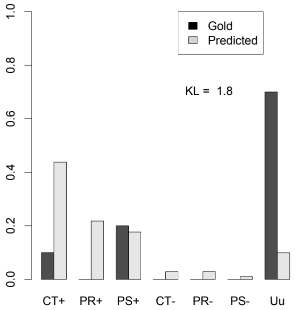


The only Uu events that the system correctly retrieved were antecedents of a conditional. For the other Uu events in (117), the system assigned CT+ or PR+. The majority of Uu events proved to be very difficult to detect automatically since complex pragmatic factors are at work, many of them only very indirectly reflected in the texts.

(117) a. Kopp's stepmother, who married Kopp's father when Kopp was in his 30s, said Thursday from her home in Irving, Texas: "I would like to see him come forward and clear his name if he's not guilty, and if he's guilty, to contact a priest and make his amends with society, face what he **did**."

Normalization: Kopp did something

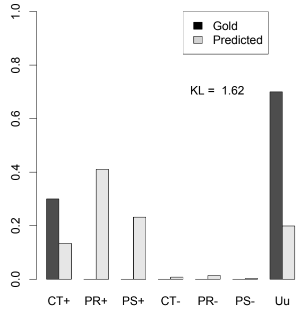
Annotations: Uu: 7, PS+: 2, CT+: 1



b. Indeed, one particularly virulent anti-abortion Web site lists the names of doctors it says perform abortions, or "crimes against humanity," with a code indicating whether they are "**working**," "wounded" or a "fatality."

Normalization: doctors are working

Annotations: Uu:7, CT+: 3

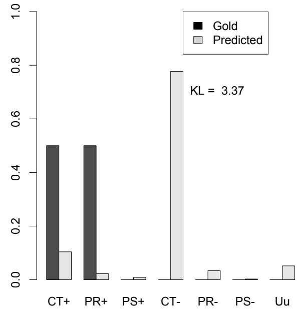


It is also instructive to look at the examples for which there is a large KL-divergence between our model's predicted distribution and the annotation distribution. Very often, this is simply the result of a divergence between the predicted and actual majority label, as discussed above. However, examples like (118) are more interesting in this regard: these are cases where there was no majority label, as in (118a), or where the model guessed the correct majority label but failed to capture other aspects of the distribution, as in (118b) and (118c).

(118) a. On Tuesday, the National Abortion and Reproductive Rights Action League plans to hold a news **conference** to screen a television advertisement made last week, before Slepian died, featuring Emily Lyons, a nurse who was badly wounded earlier this year in the bombing of an abortion clinic in Alabama.

Normalization: there will be a news conference to screen a television advertisement

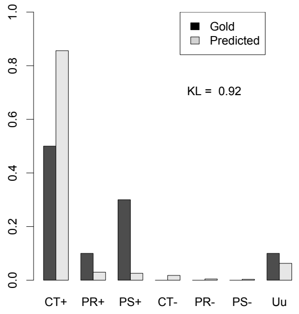
Annotations: CT+: 5, PR+: 5



b. Vacco's campaign manager, Matt Behrmann, said in a statement that Spitzer had "sunk to a new and despicable low by **attempting** to capitalize on the murder of a physician in order to garner votes."

Normalization: Spitzer had attempted to capitalize on the murder of a physician in order to garner votes

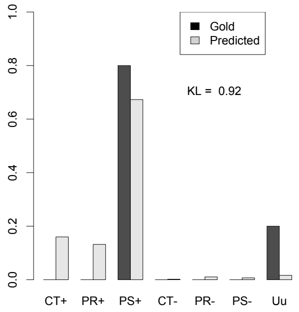
Annotations: CT+: 5, PR+: 1, PS+: 3, Uu: 1



c. Since there is no federal homicide statute as such, the federal officials said Kopp could be **charged** under the recent Freedom of Access to Clinic Entrances Act, which provides for a sentence of up to life imprisonment for someone convicted of physical assaults or threats against abortion providers.

Normalization: Kopp will be charged under the recent Freedom of Access to Clinic Entrances Act

Annotations: PS+: 8, Uu: 2

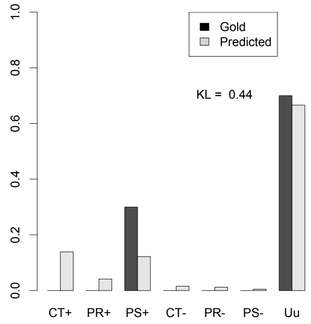


In (118a), the classifier is confused by an ambiguity: it treats *hold* as a kind of negation, which leads the system to assign a 0.78 probability to CT-. In (118b), there are no features indicating possibility, but a number of SAY-related features are present, which leads to a very strong bias for CT+ (0.86) and a corresponding failure to model the rest of the distribution properly. In (118c), the classifier correctly assigns most probability to PS+, but the rest of the probability mass is distributed between CT+ and PR+. This is another manifestation of the problem, noted above, that we have very few strong indicators of Uu. The exception to that is conditional antecedents. As a result, the system does well with cases like 119, where the event is in a conditional; the classifier assigns 70% of the probability to Uu and 0.15 to PS+.

(119) On Monday, Spitzer called for Vacco to revive that unit immediately, vowing that he would do so on his first day in office if **elected**.

Normalization: Spitzer will be elected

Annotations: Uu: 7, PS+: 3



Overall the system assigns incorrect veridicality distributions in part because it misses explicit linguistic markers of veridicality, but also because contextual and pragmatic factors cannot be fully captured. This is instructive, though, and serves to further support the central thesis that veridicality judgments are not purely lexical, but rather involve complex pragmatic reasoning.

5.8 Discussion

My central goal for this chapter was to explore veridicality judgments at the level of pragmatic meaning. To do this, I extended FactBank (Saurí & Pustejovsky 2009) with veridicality annotations that are informed by context and world knowledge (section 5.3). While the original FactBank annotations and the ones I collected are similar in many ways, their differences highlight areas in which pragmatic factors play a leading role in shaping speakers' judgments (section 5.4.1). In addition, because each one of our sentences was judged by ten annotators, I actually have annotation *distributions*, which allow me to identify areas of uncertainty in veridicality assessment (section 5.4.2). This uncertainty is so pervasive that the problem itself is better modeled as one of predicting a distribution over veridicality categories, rather than trying to predict a single label as it is usually done. The predictive model I developed (section 5.5) is true to this intuition, since it trains on and predicts distributions. All the features of the model, even the basic lexical ones, show the influence of interacting pragmatic factors (section 5.6). Although automatically assigning veridicality judgments that correspond to readers' intuitions when pragmatic factors are allowed to play a role is challenging, the classifier developed here shows that it can be done effectively using a relatively simple feature set.

These findings resonate with the notion of contradiction I put forth in chapter 4, where I emphasized that a definition of contradiction suitable to NLP tasks needs to draw on how people interpret utterances naturalistically. A similar idea of "common-sense" understanding of language lies behind the notion of entailment used in the RTE challenges (Dagan *et al.* 2006), where the goal is to determine, for each pair of sentences $\langle T, H\rangle$, whether T (the *text*) justifies H (the *hypothesis*). Entailments and contradictions are thus not calculated over just the information contained in the sentence pairs, as a more classical logical approach would have it, but rather over the full utterance meaning. As a result, they are imbued with all the uncertainty of pragmatic meaning. This is strongly reminiscent of the distinction I make between semantic and pragmatic veridicality. For example, as a purely semantic fact, *might*(S) is non-veridical with regard to S. However, depending on the nature of S, the nature of the source, the context, and countless other factors, one might nonetheless infer S. This is one of the central lessons of the new annotations presented in this chapter.

In an important sense, I have been conservative in bringing semantics and pragmatics together, because I do not challenge the basic veridicality categorizations that come from linguistic and logical work on this topic. Rather, I showed that those semantic judgments are often enriched pragmatically – for example, from uncertainty to one of the positive or negative categories, or from PS to PR or even CT. There is, however, evidence suggesting that we should be even more radically pragmatic (Searle 1978; Travis 1996), by dropping the notion that lexical items can even be reliably classified once and for all. For example, lexical theories generally agree that *know* is veridical with respect to its sentential complement, and the vast majority of its uses seem to support that claim. There are exceptions, though, as in (120) (see also Beaver (2010); Hazlett (2010)):

(120) a. But I guess when you know something terribly important that the entire world thinks is hooey, it gets harder and harder to let it go. http://www.washingtonmonthly.com/features/2009/0905.ohare.html [The author is the target of a conspiracy theory, and this is a description of the conspiracy theorist's mental state.]

b. For the first time in history, the U.S. has gone to war with an Arab and Muslim nation, and we know a peaceful solution **was in reach**.

c. Let me tell you something, when it comes to finishing the fight, Rocky and I have a lot in common. I never quit, I never give up, and I know that we're going to **make it together**.
– Hillary Clinton, September 1, 2008.

d. "That woman who knew I **had dyslexia** – I never interviewed her."
– George W. Bush (The New York Times, September 16, 2000. Quoted by Miler (2001).)

All of these examples seem to use *know* to report emphatically held belief, a much weaker sense than a factive lexical semantics would predict. Example (120d) is the most striking of the group, because it seems to be pragmatically non-veridical: the continuation is Bush's evidence that the referent of *that woman* could not possibly be in a position to determine whether he is dyslexic. In this chapter, I have shown the importance of a pragmatically-informed perspective on veridicality in natural language, and demonstrated that automatic veridicality assignment is feasible. The examples in (120) further emphasize the critical role of pragmatics in veridicality assessment.

One key component of veridicality judgment that I left out in this study is the text provenance. The data did not allow me to examine its impact since I did not have enough variation in the provenance. All FactBank sentences are from newspaper and newswire text such as the Wall Street Journal, the Associated Press, and the New York Times. However, the trustworthiness of the document provenance can affect veridicality judgments: people might have different reactions reading a sentence in the New York Times or in a random blog on the web. Online data presents a difficult case. Its provenance can affect credibility. The presence of opinion spam is a big issue: fictitious reviews, written to sound authentic, have the purpose to deceive the reader. Work on online text credibility has often focused on how external features, such as the text topic, the user name or the user image, have an impact on credibility perception (i.a., Morris *et al.* 2012). There has however been some work incorporating linguistic features. Ott *et al.* (2011) developed a corpus of 400 truthful and 400 gold-standard deceptive reviews. They show that humans are very poor judges of deception (∼60% accuracy), and build an effective system to detect deceptive reviews. Their classifier uses bigrams as well as psychological cues from the *Linguistic Inquiry and Word Count* software (Pennebaker *et al.* 2007). It achieves 90% accuracy. The psychological cues to deception only give a very small boost in performance. Other work showed that n-gram features perform well in deception detection (i.a., Jindal & Liu 2008; Mihalcea & Strapparava 2009). Ott *et al.* (2011) also emphasized that parts-of-speech are very good indicators of fake vs. truthful reviews (the classifier based on part-of-speech alone yields 73% accuracy). It would be interesting to conduct a deeper analysis of the linguistic features present in fake and real reviews, and examine how these tie in with veridicality.

Chapter 6

Conclusion

In this dissertation, I have embarked on building models for automatically handling context. By and large, language understanding in computational linguistics has focused on the literal meaning of sentences. However, to reach real natural language understanding, computational linguistics needs to capture the way humans understand and reason about language. We understand not only the words in a text, but also what is conveyed by language beyond the literal meaning of the words. We reason from context. I therefore aimed at developing computational models that capture *pragmatic meaning*, the kind of information that a reader will reliably extract from an utterance taking context into account. Such a goal was considered unattainable until recently. I found, however, that some aspects of handling context are easier than expected. There are phenomena systematic enough to be modeled computationally, and I show that the current era of rapidly expanding user-generated web content offers situated language which allows us to successfully capture pragmatic meaning.

I focus on three phenomena for which humans readily make inferences: automatically learning orderings of scalar modifiers that allow the interpretation of indirect answers to *yes/no* questions, automatically detecting conflicting information in two pieces of text, and determining the role of context and world knowledge in judging the veridicality of events. The methods I propose to solve these different issues exploit the modern data-rich world we now have access to: I use distributions in large amounts of text from the Web, as well as annotations gathered via crowdsourcing techniques. One methodological point I want to stress is that the work in my dissertation is grounded in real data. I believe, that when targeting pragmatic meaning, it is essential to analyze naturally-occurring examples, and not constructed ones, to get the full flavor of how humans use language.

My work emphasizes the effectiveness of crowdsourcing techniques for tapping into pragmatic meaning. Common-sense intuitions are the primary data for pragmatic meaning. I show that it is possible to create crowdsourcing experiments which involve naturally-occurring tasks appealing to common-sense intuitions. The crowdsourcing experiments I developed led to reliable annotations. These annotations confirmed that uncertainty is inherent when embracing pragmatics. Even though the level of meaning I focus on is the one of systematic inference based on general expectations about how language is normally used, there is some inherent variability in how people interpret utterances. Such variability should not be ignored. To adequately model pragmatic meaning, it is necessary to allow for uncertainty. I suggest that probabilistic models are well-suited to dealing with such uncertainty. Thus the three models proposed in this dissertation all have an intrinsic probabilistic component.

I also emphasize that humans do not always follow a strictly logical reasoning in their inferences. Perhaps the most illustrative example comes from the chapter on contradiction detection. If one wants to retrieve information which humans see as contradictory, it is not helpful to take a logical definition of contradiction. Rather than defining two pieces of text as contradictory if there is no possible world in which both are true, we need to define pieces of text as contradictory if they are extremely unlikely to be considered true simultaneously. Such a definition departs from the traditional view in formal semantics, but better fits people's intuitions of what a contradiction is. To capture pragmatic meaning, we need to model common-sense reasoning, which is not synonymous with logical reasoning.

I also show that when focusing on pragmatic meaning, context might shift lexical meanings. To accurately integrate inferences generated from context, we need to be able to model more than just the words. In the first case study targeting the interpretation of indirect answers to *yes/no* questions involving gradable adjectives, some pairs of modifiers found in the question and in the answer are opposites, even though they are not prototypical opposites. But the context of use renders an opposition. As a consequence, such pairs are often absent from canonical antonym lists, dictionaries or thesauri. The technique I propose, based on large amounts of situated language, finds such opposites. The impact of context on lexical meanings is also striking in the third case study aiming at assigning veridicality judgment to events. I show that when people are allowed to use context and world knowledge in judging whether they view an event as happening or not in the world, their intuitions do not always align with the predictions of lexicalist theories. Lexicalist theories always assign one unique veridicality value to a word (given the syntactic construction it is in). Veridicality values are thus deterministic according to lexical theories. However, I show that, for a given word, different contexts (which do not differ in their syntactic constructions) might yield different veridicality values. Lexical theories, by restricting information to come from the words only, offer a way to avoid reasoning with world knowledge. However, avoiding reasoning with world knowledge will not let us grasp the way humans use and understand language.

Overall I suggest that semantic judgments are often enriched pragmatically, and that such pragmatic enrichment needs to be modeled to achieve real natural language understanding. My dissertation shows how some aspects of the ways people enrich the meaning of utterances in relation to their context can be captured automatically. I have of course not solved the entire problem of natural language understanding, but I have highlighted the importance of capturing pragmatic meaning to achieve such understanding and I have provided some first steps towards reasoning from context. I suggested an approach to target pragmatic meaning, and proved its feasibility in three case studies, each one focusing on a different type of inference. The method I propose makes up for the drawbacks of earlier approaches to language understanding. Instead of allowing only certain and categorical inference, I build models which allow uncertain inference. The models are built on empirical data rather than on hand-built scripts or schemas. They are also not domain-limited. Further, they are based on common-sense reasoning, and are therefore not restricted to logical inference only.

More work lies ahead. One major issue is the acquisition of knowledge. I have demonstrated that some knowledge necessary for inference can be obtained using large corpora, but in some instances, I failed to find sufficient knowledge to allow reasoning (recall the large firm example (56) repeated here in (121), where one needs to decide whether a firm of three hundred and fifty people is considered large or not).

(121) [sw00utt/sw_0007_4171.utt]

A: Do you happen to be working for a large firm?

B: It's about three hundred and fifty people.

There is not only a need for gathering knowledge that is used in systematic inference; ideally we should also retrieve specific knowledge present in context. For instance, based on typical assumptions about work and illness, if someone is sick with the flu, people will probably infer that the person is not at work. But if the context mentions the person's character (e.g., workaholic tendencies), this information might impact the inference process and should be taken into account. Going forward, to provide deeper automatic natural language understanding, we will need to look at additional ways to gather world knowledge and model richer contextual features so as to allow them to play a role in inference.

Bibliography

Abbott, Rob, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, & Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, 2–11.

Allen, James F., & C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence* 15.143–178.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, & Regina Weinert. 1991. The HCRC map task data. *Language and Speech* 34.351–366.

Asher, Nicholas. 2000. Truth conditional discourse semantics for parentheticals. *Journal of Semantics* 17.31–50.

——, & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Austin, John. L. 1962. *How to do things with words*. Oxford: Clarendon Press.

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, & Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Barwise, Jon. 1981. Scenes and other situations. *The Journal of Philosophy* 78.369–397.

Beaver, David. 2010. Have you noticed that your belly button lint colour is related to the colour of your clothing? In *Presuppositions and Discourse: Essays Offered to Hans Kamp*, ed. by Rainer Bäuerle, Uwe Reyle, & Thomas Ede Zimmermann, 65–99. Elsevier.

Bejan, Cosmin Adrian, & Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1412–1422.

Berger, A., S. Della Pietra, & V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22.39–71.

Björne, Jari, & Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, 183–191.

Blackburn, Patrick, & Johan Bos. 2005. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI.

Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, & Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era (NLPIX)*.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Bos, Johan, & Katja Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Oxford: Blackwell.

Buchholz, Sabine, & Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164.

Carroll, John, Guido Minnen, & Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC)*.

Carston, Robyn. 1988. Implicature, explicature and truth-theoretic semantics. In *Mental Representations: The Interface between Language and Reality*, ed. by R. Kempson, 155–181, Cambridge. Cambridge University Press.

Chambers, Nathanael, 2011. *Inducing Event Schemas and their Participants from Unlabeled Text*. Department of Computer Science, Stanford University dissertation.

——, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, & Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 165–170.

Chapman, Siobhan. 2005. *Paul Grice: Philosopher and Linguist*. Palgrave Macmillan.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper, & Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* 34.301–310.

Chen, Lin, & Barbara Di Eugenio. 2012. Co-reference via pointing and haptics in multi-modal dialogues. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 523–527.

Chevalier, Judith A., & Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43.345–354.

Chklovski, Timothy, & Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 33–40.

Choi, Yoonjung, Youngho Kim, & Sung-Hyon Myaeng. 2009. Domain-specific sentiment analysis using contextual feature generation. In *TSA'09, 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, 37–44.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague/Paris: Mouton.

Clark, Herbert H. 1979. Responding to indirect speech acts. *Cognitive psychology* 11.430–477.

—— 1996. *Using language*. Cambridge: Cambridge University Press.

——, & Michael F. Schober. 1992. Asking questions and influencing answers. In *Questions about questions*, ed. by Judith M. Tanur, 15–48, New York. Russell Sage Foundation.

Clegg, Andrew B., & Adrian J. Shepherd. 2007. Evaluating and integrating treebank parsers on a biomedical corpus. *BMC Bioinformatics* 8.

Clifton, Charles Jr., & Lyn Frazier. 2010. Imperfect ellipsis: Antecedents beyond syntax? *Syntax* 13.279–297.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20.37–46.

Collins, Michael, 1999. *Head-driven statistical models for natural language parsing*. University of Pennsylvania dissertation.

Condoravdi, Cleo, Dick Crouch, Valeria de Pavia, Reinhard Stolle, & Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning - Volume 9*, 38–45.

Crammer, Koby, & Yoram Singer. 2001. Ultraconservative online algorithms for multiclass problems. In *Computational Learning Theory*, 99–115. Springer.

Crouch, Richard, Lauri Karttunen, & Annie Zaenen, 2006. Circumscribing is not excluding: A reply to Manning. Ms., Palo Alto Research Center.

Cruse, Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

——. 2011. *Meaning in Language. An Introduction to Semantics and Pragmatics*. New York: Oxford University Press.

——, & Pagona Togia. 1995. Towards a cognitive model of antonymy. *Lexicology* 1.113–141.

Dagan, Ido, Oren Glickman, & Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Lecture Notes in Computer Science*, ed. by J. Quinonero-Candela, I. Dagan, B. Magnini, & F. d'Alché-Buc, volume 3944, 177–190. Springer-Verlag.

de Marneffe, Marie-Catherine, Trond Grenager, Bill MacCartney, Daniel Cer, Daniel Ramage, Chloé Kiddon, & Christopher D. Manning. 2007. Aligning semantic graphs for textual inference and machine reading. In *AAAI Spring Symposium at Stanford*.

de Marneffe, Marie-Catherine, Scott Grimm, & Christopher Potts. 2009. Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 136–143.

de Marneffe, Marie-Catherine, Bill MacCartney, & Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 449–454.

de Marneffe, Marie-Catherine, & Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 1–8.

de Marneffe, Marie-Catherine, Christopher D. Manning, & Christopher Potts. 2010. Was it good? It was provocative. learning the meaning of scalar adjectives. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics (ACL 2010)*, 167–176.

de Marneffe, Marie-Catherine, Christopher D. Manning, & Christopher Potts. 2011a. Veridicality and utterance understanding. In *Proceedings of the 2011 Fifth IEEE International Conference on Semantic Computing*, 430–437.

de Marneffe, Marie-Catherine, Christopher D. Manning, & Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38.301–333.

de Marneffe, Marie-Catherine, Anna N. Rafferty, & Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, 1039–1047.

——, Anna R. Rafferty, & Christopher D. Manning. 2011b. Identifying conflicting information in texts. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, ed. by Joseph Olive, Caitlin Christianson, & John McCary, New York, NY. Springer.

Diab, Mona, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, & Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, 68–73.

Dolan, Bill, Chris Brockett, & Chris Quirk, 2005. Microsoft Research Paraphrase Corpus. http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/.

Drew, Paul. 1992. Contested evidence in courtroom cross-examination: The case of a trial for rape. In *Talk at work: Interaction in institutional settings*, ed. by Paul Drew & John Heritage, 470–520, New York. Cambridge University Press.

El Maarouf, Ismaïl, & Jeanne Villaneau. 2012. A French fairy tale corpus syntactically and semantically annotated. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Elkin, Peter L., Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom, & Dietlind L. Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making* 5.

Erkan, Gunes, Arzucan Ozgur, & Dragomir R. Radev. 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 228–237.

Fahrni, Angela, & Manfred Klenner. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. *Computational Linguistics* 2.60–63.

Farkas, Richárd, Veronika Vincze, György Móra, János Csirik, & György Szarvas. 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*, 1–12.

Fauconnier, Gilles. 1975. Pragmatic scales and logical structure. *Linguistic Inquiry* 6.353–375.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database*. MIT Press.

Finlay, Stephen. 2009. Oughts and ends. *Philosophical Studies* 143.315–340.

Firbas, Jan. 1971. On the concept of communicative dynamism in the theory of functional sentence perspective. *Brno Studies in English* 7.23–47.

Fleiss, Joseph I. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76.378–382.

Frazier, Lyn, & Charles Jr. Clifton. 2005. The syntax-discourse divide: Processing ellipsis. *Syntax* 8.121–174.

Fundel, Katrin, Robert Küffner, & Ralf Zimmer. 2007. RelEx-relation extraction using dependency parse trees. *Bioinformatics* 23.

Garten, Yael, 2010. *Text mining of the scientific literature to identify pharmacogenomic interactions*. Department of Biomedical Informatics, Stanford University dissertation.

Genzel, Dmitriy. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 376–384.

Giampiccolo, Danilo, Ido Dagan, Bernardo Magnini, & Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 1–9.

Giannakidou, Anastasia. 1994. The semantic licensing of NPIs and the Modern Greek subjunctive. In *Language and Cognition 4, Yearbook of the Research Group for Theoretical and Experimental Linguistics*, 55–68. University of Groningen.

——. 1995. Weak and strong licensing: Evidence from Greek. In *Studies in Greek Syntax*, ed. by Artemis Alexiadou, Geoffrey Horrocks, & Melita Stavrou, 113–133. Dordrecht: Kluwer.

——. 1999. Affective dependencies. *Linguistics and Philosophy* 22.367–421.

——. 2001. The meaning of free choice. *Linguistics and Philosophy* 24.659–735.

Glinos, Demetrios G. 2010. System description for SAIC entry at RTE-6. In *Proceedings of the Text Analysis Conference (TAC)*.

Godfrey, John J., Edward C. Holliman, & Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing*.

Green, Nancy, & Sandra Carberry. 1994. A hybrid reasoning model for indirect answers. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 58–65.

——, & ——. 1999. Interpreting and generating indirect answers. *Computational Linguistics* 25.389–435.

Grice, H. Paul. 1975. Logic and conversation. In *Syntax and semantics*, ed. by P. Cole & J. Morgan, volume 3. Cambridge, MA: Academic Press.

Groenendijk, Jeroen, & Martin Stokhof, 1984. *Studies in the semantics of questions and the pragmatics of answers*. University of Amsterdam dissertation.

Haghighi, Aria, & Dan Klein. 2010. An entity-level approach to information extraction. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics, Short Papers*, 291–295.

Harabagiu, Sanda, Andrew Hickl, & Finley Lacatusu. 2006. Negation, contrast, and contradiction in text processing. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, 755–762.

Harris, Jesse A., & Christopher Potts. 2009. Perspective-shifting with appositives and expressives. *Linguistics and Philosophy* 32.523–552.

Hassan, Ahmed, Vahed Qazvinian, & Dragomir Radev. 2010. What's with the attitude?: Identifying sentences with attitude in online discussions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, 1245–1255.

Haverinen, Katri, Filip Ginter, Timo Viljanen, Veronika Laippala, & Tapio Salakoski. 2010a. Dependency-based propbanking of clinical Finnish. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, 137–141.

——, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, & Tapio Salakoski. 2010b. Treebanking Finnish. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT).*

Hays, David G. 1964. Dependency theory: A formalism and some observations. *Language* 40.511–525.

——, & Theodore W. Ziehe. 1960. *Studies in Machine Translation - 10: Russian Sentence-Structure Determination*. Research Memorandum RM-2538. Santa Monica, California: The RAND Corporation.

Hazlett, Allan. 2010. The myth of factive verbs. *Philosophy and Phenomenological Research* 80.497–522.

Hirschberg, Julia B., 1985. *A Theory of Scalar Implicature*. University of Pennsylvania dissertation.

Hobby, Jonathan L., Brian D.M. Tom, C. Todd, Philip W.P. Bearcroft, & Adrian K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology* 73.999–1001.

Hockey, Beth Ann, Deborah Rossen-Knill, Beverly Spejewski, Matthew Stone, & Stephen Isard. 1997. Can you predict answers to Y/N questions? Yes, No and Stuff. In *Proceedings of Eurospeech 1997*.

HORN, LAURENCE R, 1972. *On the Semantic Properties of Logical Operators in English*. Los Angeles: UCLA dissertation.

HU, NAN, PAUL A. PAVLOU, & JENNIFER ZHANG. 2006. Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of Electronic Commerce (EC)*, 324–330.

HUANG, YANG, & HENRY J. LOWE. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association* 14.304–311.

JINDAL, NITIN, & BING LIU. 2008. Opinion spam and analysis. In *Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008)*, 219–230.

JOSHI, MAHESH, DIPANJAN DAS, KEVIN GIMPEL, & NOAH A. SMITH. 2010. Movie reviews and revenues: an experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 293–296.

JURAFSKY, DANIEL, ELIZABETH SHRIBERG, & DEBRA BIASCA. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science.

KAMP, HANS. 1975. Two theories about adjectives. In *Formal semantics of natural language*, ed. by E. L. Keenan, 123–155, Cambridge. Cambridge University.

——, & BARBARA H. PARTEE. 1995. Prototype theory and compositionality. *Cognition* 57.129–191.

KARTTUNEN, LAURI. 1973. Presuppositions and compound sentences. *Linguistic Inquiry* 4.169–193.

——, & Annie Zaenen. 2005. Veridicity. In *Annotating, Extracting and Reasoning about Time and Events*, ed. by Graham Katz, James Pustejovsky, & Frank Schilder, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30.1–45.

——, & Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language* 81.345–381.

Kessler, Jason S. 2008. Polling the blogosphere: a rule-based approach to belief classification. In *International Conference on Weblogs and Social Media*.

Kim, Hyun Duk, & ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and Knowledge Management*, 385–394.

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, & Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, 1–9.

——, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, & Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, 1–6.

King, Tracy H., Richard Crouch, Stefan Riezler, Mary Dalrymple, & Ronald Kaplan. 2003. The PARC 700 dependency bank. In *4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.

Kiparsky, Paul, & Carol Kiparsky. 1970. Facts. In *Progress in linguistics*, ed. by M. Bierwisch & K. E. Heidolph, The Hague, Paris. Mouton.

Klein, Dan, & Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association of Computational Linguistics*, 423–430.

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4.1–45.

Klüwer, Tina, Hans Uszkoreit, & Feiyu Xu. 2010. Using syntactic and semantic based relations for dialogue act recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 570–578.

Kouylekov, Milen, Yashar Mehdad, Matteo Negri, & Elena Cabrio. 2010. FBK participation in RTE6: Main and KBP validation task. In *Proceedings of the Text Analysis Conference (TAC)*.

Kratzer, Angelika. 1981. The notional category of modality. In *Words, worlds, and contexts. New approaches in Word Semantics*, ed. by Hans-Jürgen Eikmeyer & Hannes Rieser, 38–74. Berlin: de Grutyer.

——. 1991. Modality. In *Semantics: An international handbook of contemporary research*, ed. by Arnim von Stechow & Dieter Wunderlich. Berlin: de Grutyer.

Kübler, Sandra, Ryan McDonald, & Joakim Nivre. 2009. Dependency parsing. In *Synthesis Lectures on Human Language Technologies*, volume 1, 1–127.

Landeghem, Sofie Van, Jari Björne, Thomas Abeel, Bernard De Baets, Tapio Salakoski, & Yves Van de Peer. 2012. Semantically linking molecular entities in literature through entity relationships. *BMC Bioinformatics* 13.

Lassiter, Daniel, 2011. *Measurement and modality: the scalar basis of modal semantics*. Department of Linguistics, New York University dissertation.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, & Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 591–601.

LEE, HEEYOUNG, MARTA RECASENS, ANGEL CHANG, MIHAI SURDEANU, & DAN JURAFSKY. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 489–500.

LEHRER, ADRIENNE J., & KEITH LEHRER. 1982. Antonymy. *Linguistics and Philosophy* 5.

LEVINSON, STEPHEN C. 1995. Three levels of meaning: Essays in honor of Sir John Lyons. In *Grammar and Meaning*, ed. by Frank R. Palmer, 90–115. Cambridge University Press.

—— 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

LEVY, ROGER, & GALEN ANDREW. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2231–2234. http://www-nlp.stanford.edu/software/tregex.shtml.

——, & CHRISTOPHER D. MANNING. 2004. Deep dependencies from context-free statistical parsers: correcting the surface dependency approximation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 327–334.

LYONS, JOHN. 1977. *Semantics*. Cambridge: Cambridge University Press.

MACCARTNEY, BILL, TROND GRENAGER, MARIE-CATHERINE DE MARNEFFE, DANIEL CER, & CHRISTOPHER D. MANNING. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 41–48.

Malakasiotis, Prodromos. 2009. AUEB at TAC 2009. In *Proceedings of the Text Analysis Conference (TAC)*.

Mann, William C., & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8.167–182.

Manning, Christopher D., 2006. Local textual inference: it's hard to circumscribe, but you know it when you see it – and NLP needs it. Ms., Stanford University.

——, & Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Tutorial at HLT-NAACL 2003 and ACL 2003*. http://nlp.stanford.edu/software/classifier.shtml.

Marcu, Daniel, & Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 368–375.

McClosky, David, & Christopher D. Manning. 2012. Learning constraints for consistent timeline extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 873–882.

Mehdad, Yashar, Matteo Negri, Elena Cabrio, Milen Kouylekov, & Bernardo Magnini. 2009. Using lexical resources in a distance-based approach to rte. In *Proceedings of the Text Analysis Conference (TAC)*.

Mihalcea, Rada, & Carlo Strapparava. 2009. The lie detector: explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 309–312.

Miler, Mark Crispin. 2001. *The Bush Dyslexicon*. W. W. Norton and Company.

Mohammad, Saif, Bonnie Dorr, & Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning (EMNLP-2008), 982–991.

Mohammad, Saif M., Bonnie J. Dorr, Graeme Hirst, & Peter D. Turney. 2011. Measuring degrees of semantic opposition. In *Technical report, National Research Council Canada, Ottawa, Canada*.

Moldovan, Dan, Christine Clark, Sanda Harabagiu, & Steven Maiorano. 2003. COGEX: A logic prover for question answering. In *Proceedings of the North American Association of Computational Linguistics (NAACL-03)*, 87–93.

Montague, Richard. 1969. On the nature of certain philosophical entities. *The Monist* 2.159–194.

Morante, Roser, & Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 21–29.

——, & Caroline Sporleder, 2010. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing.*

——, & ——. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics* 38.223–258.

Morris, Meredith Ringel, Scott Counts, Asta Roseway, Aaron Hoff, & Julia Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 441–450.

Mosteller, Frederick, & Cleo Youtz. 1990. Quantifying probabilistic expressions. *Statistical Science* 5.2–34.

Nairn, Rowan, Cleo Condoravdi, & Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of Inference in Computational*

Semantics 5, ed. by Johan Bos & Alexander Koller, 67–76. Buxton, England: ACL.

Norvig, Peter, 1986. *A unified theory of inference for text understanding*. UC Berkeley dissertation.

Ott, Myle, Yejin Choi, Claire Cardie, & Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 309–319.

Pakray, Partha, Snehasis Neogi, Pinaki Bhaskar, Soujanya Poria, Sivaji Bandyopadhyay, & Alexander Gelbukh. 2011. A textual entailment system using anaphora resolution. In *Proceedings of the Text Analysis Conference (TAC)*.

Palmer, Frank R. 1986. *Mood and modality*. Cambridge: Cambridge University Press.

Pang, Bo, & Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 271–278.

——, & ——. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2.1–135.

——, ——, & Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86.

Pasca, Marius, & Sanda Harabagiu. 2001. High performance question/answering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 366–374.

Paul, Michael J., ChengXiang Zhai, & Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 66–76.

Pennebaker, James W., Cindy K. Chung, Molly Ireland, Amy Gonzales, & Roger J. Booth, 2007. The development and psychometric properties of LIWC2007. LIWC.net.

Perlmutter, David M. (ed.) 1983. *Studies in Relational Grammar*, volume 1. Chicago, IL: University of Chicago Press.

Perrault, C. Raymond, & James F. Allen. 1980. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics* 6.167–182.

Petrov, Slav, & Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *First Workshop on Syntactic Analysis of Non-Canonical Language*.

Pomerantz, Anita M. 1984. Agreeing and disagreeing with assessment: Some features of preferred/dispreferred turn shapes. In *Structure of Social Action: Studies in Conversation Analysis*, ed. by J. M. Atkinson & J. Heritage. Cambridge University Press.

Portner, Paul. 2009. *Modality*. Oxford University Press.

Potisuk, Siripong. 2010. Typed dependency relations for syntactic analysis of Thai sentences. In *Proceedings of PACLIC 24 Pacific Asia Conference on Language, Information and Computation*.

Prabhakaran, Vinodkumar, Owen Rambow, & Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 1014–1022.

PUSTEJOVSKY, JAMES, MARC VERHAGEN, ROSER SAURÍ, JESSICA LITTMAN, ROBERT GAIZAUSKAS, GRAHAM KATZ, INDERJEET MANI, ROBERT KNIPPEN, & ANDREA SETZER. 2006. Timebank 1.2. In *Linguistic Data Consortium*, Philadelphia, PA.

PYYSALO, SAMPO, FILIP GINTER, KATRI HAVERINEN, JUHO HEIMONEN, TAPIO SALAKOSKI, & VERONIKA LAIPPALA. 2007. On the unification of syntactic annotations under the stanford dependency scheme: A case study on BioInfer and GENIA. In *Proceedings of BioNLP 2007: Biological, translational, and clinical language processing (ACL07)*, 25–32.

——, TOMOKO OHTA, & JUNICHI TSUJII. 2011. An analysis of gene/protein associations at PubMed scale. *Journal of Biomedical Semantics* 2.

RECANATI, FRANÇOIS. 2004a. *Literal meaning*. Cambridge: Cambridge University Press.

——. 2004b. Pragmatics and semantics. In *Handbook of Pragmatics*, ed. by Laurence R. Horn & Gregory Ward, 442–462. Oxford: Blackwell.

REYNA, VALERIE F. 1981. The language of possibility and probability: Effects of negation on meaning. *Memory and Cognition* 9.642–650.

RIEH, SOO YOUNG. 2010. Credibility and cognitive authority of information. In *Encyclopedia of Library and Information Sciences, 3rd Ed.*, ed. by M. Bates & M. N. Maack, 1337–1344. New York: Taylor and Francis Group, LLC.

RITTER, ALAN, DOUG DOWNEY, STEPHEN SODERLAND, & OREN ETZIONI. 2008. It's a contradiction – no, it's not: A case study using functional relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 11–20.

ROBERTS, CRAIGE. 1996. Information structure: Towards an integrated formal theory of pragmatics. In *OSU Working Papers in Linguistics*, ed. by Jae Hak

Yoon & Andreas Kathol, volume 49: Papers in Semantics, 91–136. Columbus, OH: The Ohio State University Department of Linguistics. Revised 1998.

Rooryck, Johan. 2001. Evidentiality, Part I. *Glot International* 5.3–11.

van Rooy, Robert. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy* 26.727–763.

Ross, John Robert. 1973. Slifting. In *The Formal Analysis of Natural Languages*, ed. by Maurice Gross, Morris Halle, & Marcel-Paul Schützenberger, 133–169. The Hague: Mouton de Gruyter.

Rubin, Victoria L. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 141–144.

——, Elizabeth D. Liddy, & Noriko Kando. 2005. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*, ed. by J. G. Shanahan, Y. Qu, & J. Wiebe. Springer-Verlag New York, Inc.

Sæbo, Kjell Johan. 2001. Necessary conditions in a natural language. In *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, ed. by Caroline Féry & Wolfgang Sternefeld, 427–449. Berlin: Akademie Verlag.

Sanchez-Graillet, Olivia, & Massimo Poesio. 2007. Discovering contradiction protein-protein interactions in text. In *Proceedings of BioNLP 2007: Biological, translational, and clinical language processing (ACL07)*, 195–196.

Sassoon, Galit M., 2010. A typology of multidimensional predicates. Ms., ILLC, University of Amsterdam, Amsterdam, The Netherlands.

Saurí, Roser, 2008. FactBank 1.0 annotation guidelines. Ms., Brandeis University.

——, & James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation* 43.227–268.

Schank, Roger C., & Robert P. Abelson. 1977. *Scripts, plans, goals and understanding. An inquiry into human knowledge structures*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schegloff, Emanuel A., Gail Jefferson, & Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53.361–382.

Searle, John R. 1978. Literal meaning. *Erkenntnis* 13.207–224.

Seraji, Mojgan, Beáta Megyesi, & Joakim Nivre. 2012. A basic language resource kit for Persian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Sheng, Victor S., Foster Provost, & Panagiotis G. Ipeirotis. 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*, 614–622.

Shivhare, Himanshu, Parul Nath, & Anusha Jain. 2010. Semi cognitive approach to RTE 6 - Using FrameNet for semantic clustering. In *Proceedings of the Text Analysis Conference (TAC)*.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117.1034–1056.

Sing, Thoudam Doren, & Sivaji Bandyopadhyay. 2010. Statistical machine translation of English – Manipuri using morpho-syntactic and semantic information. In *Proceedings of the Association for Machine Translation in the Americas (AMTA 2010)*.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, & Andrew Y. Ng. 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*, 254–263.

Stenström, Anna-Brita. 1984. Questions and responses in English conversation. In *Lund Studies in English 68*, ed. by Claes Schaar & Jan Svartvik, Malmö Sweden. CWK Gleerup.

Szarvas, György. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of 46th Meeting of the Association for Computational Linguistics*, 281–289.

——, Veronika Vincze, Richárd Farkas, & János Csirik. 2008. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 38–45.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Travis, Charles. 1996. Meaning's role in truth. *Mind* 105.451–466.

Tversky, Amos, & Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science, New Series* 211.453–458.

Urbain, Jay, Nazli Goharian, & Ophir Frieder. 2007. IIT TREC 2007 genomics track: Using concept-based semantics in context for genomics literature passage retrieval. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*.

Vanderwende, Lucy, Arul Menezes, & Rion Snow. 2006. Microsoft Research at RTE-2: Syntactic contributions in the entailment task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

von Fintel, Kai. 2006. Modality and language. In *Encyclopedia of Philosophy*, ed. by Donald. M. Borchert. Detroit, MI: MacMillan Reference.

——, & Sabine Iatridou. 2008. How to say *ought* in foreign: The composition of weak necessity modals. In *Studies in Natural Language and Linguistic Theory*, ed. by Jacqueline Guéron & Jacqueline Lecarme, volume 75, 115–141. Berlin: Springer.

Voorhees, Ellen. 2008. Contradictions and justifications: Extensions to the textual entailment task. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 63–71.

Wang, Rui, & Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 784–792.

Warren, David H., & Fernando C. N. Pereira. 1982. An efficient easily adaptable system for interpreting natural language queries. *American Journal of Computational Linguistics* 8.110–122.

Wells, Rulon. 1947. Immediate constituents. *Language* 23.81–117.

Wiebe, Janyce. 1994. Tracking point of view in narrative. *Computational Linguistics* 20.233–287.

——, & Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 38.165–210.

——, Theresa Wilson, Rebecca Bruce, Matthew Bell, & Melanie Martin. 2004. Learning subjective language. *Computational Linguistics* 30.277–308.

Wierzbicka, Anna. 1987. The semantics of modality. *Folia Linguistica* 21.25–43.

Wilson, Theresa, 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Pittsburgh, PA: University of Pittsburgh dissertation.

——, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, & Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Demonstration Description in Conference on Empirical Methods in Natural Language Processing*, 34–35.

Wu, Fei, & Daniel S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 118–127.

Xu, Peng, Jaeho Kang, Michael Ringgaard, & Franz Och. 2009. Using a dependency parser to improve SMT for subject-object-verb languages. In *NAACL 2009: Proceedings of Human Language Technologies, The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 245–253.

Zaenen, Annie. forthcoming. Do give a penny for their thoughts. To appear in *Natural Language Engineering.*

——, Lauri Karttunen, & Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 31–36.

Zeevat, Henk. 1994. Questions and exhaustivity in update semantics. In *Proceedings of the International Workshop on Computational Semantics*, ed. by Harry Bunt, Reinhard Muskens, & Gerrit Rentier, 211–221.

Zouaq, Amal, Michel Gagnon, & Benoît Ozell. 2010. Semantic analysis using dependency-based grammars and upper-level ontologies. *International Journal of Computational Linguistics and Applications* 1.85–101.

Zwarts, Frans. 1995. Nonveridical contexts. *Linguistic Analysis* 25.286–312.