

BAR ILAN UNIVERSITY

APPLIED TEXTUAL ENTAILMENT

Oren Glickman

Department of Computer Science

Ph.D. Thesis

Submitted to the Senate of Bar Ilan University

Ramat Gan, Israel June 2006

This work was carried out under the supervision of Dr. Ido Dagan and Prof. Moshe Koppel (Department of Computer Science), Bar-Ilan University.

Abstract

This thesis introduces the applied notion of textual entailment as a generic empirical task that captures major semantic inferences across many applications. Textual entailment addresses semantic inference as a direct mapping between language expressions and abstracts the common semantic inferences as needed for text based Natural Language Processing applications. We define the task and describe the creation of a benchmark dataset for textual entailment along with proposed evaluation measures. This dataset was the basis for the PASCAL Recognising Textual Entailment (RTE) Challenge. We further describe how textual entailment can be approximated and modeled at the lexical level and propose a lexical reference subtask and a correspondingly derived dataset.

The thesis further proposes a general probabilistic setting that casts the applied notion of textual entailment in probabilistic terms. We suggest that the proposed setting may provide a unifying framework for modeling uncertain semantic inferences from texts. In addition, we describe two lexical models demonstrating the applicability of the probabilistic setting. Although our proposed models are relatively simple, as they do not rely on syntactic or other deeper analysis, they nevertheless achieved competitive results on the PASCAL RTE challenge.

Finally, the thesis presents a novel acquisition algorithm to identify lexical entailment relations from a single corpus focusing on the extraction of verb paraphrases. Most previous approaches detect individual paraphrase instances within a pair (or

set) of comparable corpora, each of them containing roughly the same information, and rely on the given substantial level of correspondence of such corpora. We present a novel method that successfully detects isolated paraphrase instances within a single corpus without relying on any a-priori structure and information. Our instance based approach seems to address some of the drawbacks of distributional similarity based methods, in particular by providing a consistent scoring scale across different words.

Preface

Portions of this thesis are joint work and have appeared elsewhere.

Chapter 3 is based on “The PASCAL Recognising Textual Entailment Challenge” LNAI book chapter (Dagan, Glickman, and Magnini, 2005). Chapter 4 and portions of chapter 5 are adapted from “A Lexical Alignment Model for Probabilistic Textual Entailment” LNAI book chapter (Glickman, Dagan, and Koppel, 2005b). Chapter 5 is also based in part on “A Probabilistic Classification Approach for Lexical Textual Entailment,” which appeared in the proceedings of the twentieth national conference on artificial intelligence (AAAI-05) (Glickman, Dagan, and Koppel, 2005a). Chapter 6 is based on “Acquiring lexical paraphrases from a single corpus,” which appeared as a book chapter in *Recent Advances in Natural Language Processing III* (Glickman and Dagan, 2004).

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors’ views.

Acknowledgements

First, I would like to thank my primary advisor, Ido Dagan, for being such a wonderful advisor. Through his support, care, and patience, he has transformed a struggling graduate student into an experienced researcher. His insight and ideas formed the foundation of this dissertation as much as mine did, and his guidance and care helped me get over various hurdles during my graduate years. Ido was a role model in many aspects and had prepared me for academic life and it has been a great experience and opportunity to work with him.

I also owe a debt of gratitude to others on my committee. I thank Moshe Koppel for his openness, his breadth of knowledge of the literature and many fruitful discussions. Jacob Goldberger, though not officially on my committee, was of significant contribution to this work. I thank Jacob for sharing with me his in depth understanding of the exciting world of generative models and Expectation Maximization.

I wish also to thank colleagues and friends at Bar Ilan. In particular the former PhD students Yuval Krymolowski and Zvika Marx as well as Idan Szpektor and Roy Bar-Haim who are currently still in the pipeline. I'd like to thank them for all their help, advice and fruitful discussions on life and science. It was great sharing an office and going out for lunches with them all.

I'd especially like to thank the other members of the PASCAL pump priming project: Eric Gaussier, Cyril Goutte, Samy Bengio, Mikaela Keller, Walter Daelemans and Anja Höthker for a fruitful collaboration. And in particular the XRCE members

for their great hospitality.

Special thanks go to those involved in the RTE challenge organization and Bernardo Magnini in particular – it was a great experience to jointly put up the RTE dataset and organize together the challenge and the workshop. I would also like to acknowledge the people and organizations who made sources available for the challenge and the people involved in creating and annotating the data: Danilo Giampiccolo, Tracy Kelly, Einat Barnoy, Alessandro Valin, Ruthie Mandel, and Melanie Joseph. Many thanks go also to the (too many to list) participants who contributed with their ideas and feedback.

Among our many colleagues I'd like to give special thanks to Dan Roth for fruitful discussions and advice.

I would like to thank Amir Ashkenazi with whom things I learned during my studies found their way to commercial products. I'll also like to thank Amir and Nahum Sharfman who kept nudging me to finish this chapter of my life and move on.

Finally, I thank my parents who have always encouraged me to pursue education and the extended family members for instilling in me confidence and a drive for pursuing my PhD. And for the one who has made the many hours spent seem worthwhile after all, Gali.

Contents

Abstract	iii
Preface	v
Acknowledgements	vi
Contents	xi
1 Introduction	1
1.1 Inference as a Relation Between Language Expressions	2
1.2 Addressing Uncertainty	4
1.3 Thesis Highlights and Contributions	5
2 Background	9
2.1 Application Needs	9
2.1.1 Question Answering (QA)	9
2.1.2 Information Extraction (IE)	10
2.1.3 Information Retrieval (IR)	11
2.1.4 Summarization	12
2.1.5 Generation	12
2.1.6 Automatic Machine Translation (MT) Evaluation	12

2.2	Techniques and Methods	13
2.2.1	Thesaurus-based Term Expansion	13
2.2.2	Distributional Similarity	14
2.2.3	Lexical Overlap	15
2.2.4	Paraphrase Acquisition	15
2.2.5	Mapping to Logical Form	18
2.3	Summary	19
3	The Applied Textual Entailment Recognition Task	21
3.1	Task Definition	22
3.1.1	Judgement Guidelines	23
3.1.2	Mapping to Applications	24
3.2	Creating the PASCAL Evaluation Dataset	25
3.2.1	Dataset Preparation	27
3.2.2	Application Settings	28
3.2.3	The Annotation Process	32
3.3	Evaluation Measures	33
3.4	Dataset Analysis	35
3.5	The Lexical Reference Subtask	40
3.5.1	Motivation	40
3.5.2	Dataset Creation and Annotation Process	41
3.5.3	Data Analysis	45
3.6	Discussion	45
4	A Probabilistic setting for Textual Entailment	51
4.1	Motivation	51
4.2	A Generative Probabilistic Setting	53
4.2.1	Probabilistic textual entailment definition	55

4.3	Model Properties	56
5	Lexical Models for Textual Entailment	59
5.1	Introduction	59
5.2	Alignment Model	60
5.2.1	Web-based Estimation of Lexical Entailment Probabilities . .	62
5.2.2	Corpus-based Estimation of Lexical Entailment Probabilities .	68
5.2.3	Performance on the Lexical Reference Subtask	70
5.3	Bayesian Model	71
5.3.1	Introduction	71
5.3.2	Textual Entailment as Text Classification	71
5.3.3	Initial Labeling	72
5.3.4	Naïve Bayes Refinement	72
5.3.5	Experimental Setting	73
5.3.6	Empirical Results	74
5.3.7	Running Additional EM Iterations	77
5.4	Discussion	79
5.5	Related Work	82
6	Acquiring Lexical Entailment Relations	85
6.1	Motivation	85
6.2	Algorithm	88
6.2.1	Preprocessing and Representation	88
6.2.2	Identifying Candidate Verb Instance Pairs (Filtering)	89
6.2.3	Computing the Paraphrase Score of Verb Instance Pairs	90
6.2.4	Computing Paraphrase Score for Verb Type Pairs	90
6.3	Evaluation and Analysis	91
6.3.1	Setting	91

6.3.2	Results of the Paraphrase Identification Algorithm	92
6.3.3	Comparison with (Lin & Pantel 2001)	95
6.4	Conclusions	97
7	Conclusions	99
7.1	The Textual Entailment Task and Evaluation Framework	99
7.2	The Probabilistic Setting and Derived Models	101
7.3	Learning Entailment Rules	102
	Bibliography	111

List of Tables

3.1	Examples of text-hypothesis pairs from the PASCAL Recognising Textual Entailment Challenge Dataset	26
3.2	Accuracy and cws results for the system submissions, ordered by first author. Partial coverage refers to the percentage of examples classified by the system out of the 800 test examples.	36
3.3	Lexical Reference Annotation Examples	44
3.4	examples demonstrating when lexical entailment does not correlate with entailment	46
4.1	Example of certain and uncertain inferences	52
4.2	Examples of various uncertain inferences	53
5.1	Accuracy results by task	63
5.2	Average precision results for alignment and baseline models	66
5.3	Accuracy and average precision for the various co-occurrence levels.	68
5.4	The lexical entailment probability estimation process - P_0 represents the initial labeling and P_1 the Bayesian estimation for $P(Tr_{job} = 1 t)$	75
5.5	Top scoring trigger words for <i>job</i> and $\neg job$	76
5.6	A sample from the lexical reference dataset along with the Bayesian model's score	81

6.1	Example of system output with judgments	93
6.2	Examples of instance pairs	94
6.3	Top 20 verb pairs from similarity system along with their similarity score	96

List of Figures

1.1	Extract from a Reading Comprehension Test	1
1.2	Logical Inference	3
1.3	Illustration of lexical variability and ambiguity	4
3.1	Accuracy results for RTE submissions	38
3.2	Mean and variance of accuracy by task	39
3.3	F-measure results for the RTE submissions	39
3.4	The breakdown of the lexical entailment dataset by type out of the 63% of the examples that were judged as true	43
5.1	System’s underlying alignment for example 1026 (RC). gold standard - false, system - false	64
5.2	System’s underlying alignment for example 1095 (202). gold standard - true, system - true	65
5.3	Comparison to baselines (<i>system</i> refers to our probabilistic model) . .	67
5.4	Recall-Precision on the RTE test using co-occurrence counts at various levels from the Reuters corpus	69
5.5	Lexical Reference average precision results for alignment model based on corpus co-occurrence counts at different co-occurrence levels	70
5.6	Recall Precision comparison on the RTE dataset for Reuters based alignment and bayesian models.	77

5.7	comparison of Recall Precision and average precision for Reuters based alignment and bayesian models on the lexical entailment task.	78
5.8	Average precision results on lexical entailment dataset for different vs. number of EM iterations	79
5.9	Mean error rates for the alignment and Bayesian models on the lexical reference dataset. Note that a lower mean error is better.	80
6.1	Example of extracting the lexical paraphrase ⟨separate, split⟩ from distinct stories	87
6.2	Extracted verb instances for sentence “But U.N. Secretary-General Boutros Boutros-Ghali delayed implementation of the deal after Iraqi forces attacked Kurdish rebels on August 31.”	88
6.3	Precision (y axis) recall (x axis) curves of system paraphrases by judge (verb type pairs sorted by system score)	92
6.4	Precision recall curve for our paraphrase method and LP similarity	97

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
CD	Comparable Documents
CWS	Confidence Weighted Score
EM	Expectation Maximization
IE	Information Extraction
IDF	Inverse Document Frequency
ILP	Inductive Logic Programming
IR	Information Retrieval
MT	Machine Translation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PP	ParaPhrases
RC	Reading Comprehension
RTE	Recognising Textual Entailment
QA	Question Answering

“It is not really difficult to construct a series of inferences, each dependent upon its predecessor and each simple in itself. If, after doing so, one simply knocks out all the central inferences and presents one’s audience with the starting-point and the conclusion, one may produce a startling, though perhaps a meretricious, effect.”

— Sherlock Holmes in “The Dancing Men”

Chapter 1

Introduction

Where Do Canadian Bugs Go In Winter?

Some insects (like many retired Canadians) prefer to take a vacation from winter. Monarch butterflies migrate south in large numbers in the fall. Some even travel 1800 miles (2700 km) from Canada to Mexico to escape winter’s chill.

Read the text above and decide if the following statements are true or false based on the reading:

- A) Many older Canadians like to take a vacation from winter.
- B) Insects, such as mosquitoes, can pass diseases to humans.
- C) Some butterflies fly over 2500 kilometres to their winter homes.

Figure 1.1: Extract from a Reading Comprehension Test

1.1 Semantic Inference as a Relation Between Language Expressions

Figure 1.1 is an extract from a reading comprehension test for students learning English as a second language (Contact, 2003). In these commonly used tests, a reader is asked to identify when a textual statement (which we term *hypothesis*) can be inferred from a given *text*. Identifying such a fundamental relation between texts is a complex task requiring a depth of language understanding. For this reason, such tests are considered a core Natural Language Understanding (NLU) task (Allen, 1995).

Many Natural Language Processing (NLP) applications also need to recognize when the meaning of one text can be expressed by, or inferred from, another text. Information Retrieval (IR), Question Answering (QA), Information Extraction (IE), text summarization and Machine Translation (MT) evaluation are examples of applications that need to assess this core semantic relationship between text segments. For example, given a question such as “Where do Canadian Monarch butterflies migrate to in the winter?” a QA system has to identify texts, such as in Figure 1.1, from which one can infer the hypothesized answer form “Canadian Monarch butterflies migrate to Mexico in the winter”. We term this relation between language expressions *textual entailment* and propose its recognition and modeling as a new generic application-independent task.

In the Artificial Intelligence (AI) and Linguistics communities, a sentence is said to entail another if the second is necessarily true given the first. The connection between sentences and facts is provided by semantics. The property of one fact following from other facts, is mirrored by the property of one sentence being entailed by others (see Figure 1.2, taken from (Russell and Norvig, 1995, page 158)). Consequently, reasoning is done on representations of facts.

When addressing inference over natural language expressions (rather than on some

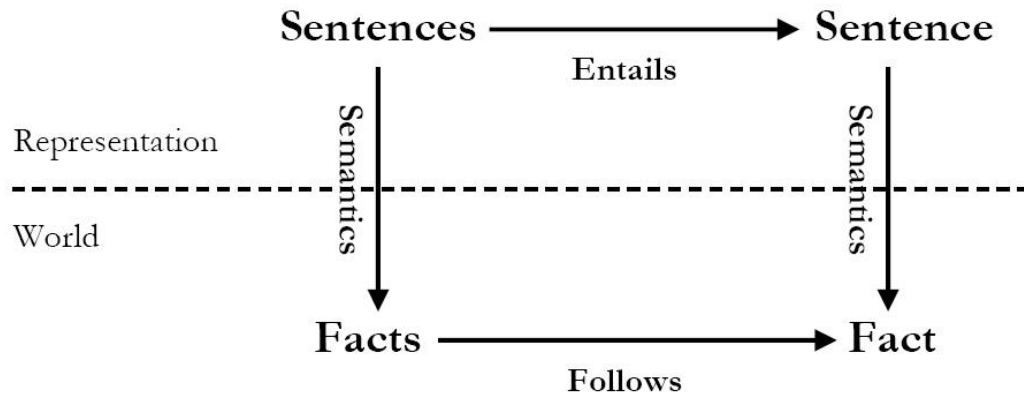


Figure 1.2: Logical Inference

other artificial meaning representation language) there are additional issues at play distinguishing it from logical inference as common in AI. When dealing with sentences in natural language, we usually refer to the more abstract notion of *meaning* rather than just facts. Human languages are extremely rich and ambiguous resulting in a many to many mapping between sentences and facts (or meanings). On the one hand an ambiguous text might represent several distinct meanings and on the other hand due to language variability a concrete meaning might be described in different ways (see Figure 1.3). Textual entailment is thus much more than just logical inference – when dealing with textual inferences one needs to account for all factors involved. For example, in figure 1.1, the truth of hypothesis C given the text is due both to *km* and *kilometres* referring to the same meaning as well as to inferences such as the relation between traveling and flying.

In this thesis we take an applicative perspective and focus on an applied notion of textual entailment – i.e. semantic inferences needed by text based NLP applications. We realize that applications deal directly with texts in natural language. For this reason, we address semantic inference as a (direct) mapping between language expressions. Leaving the interpretation into meaning representations (e.g. explicitly

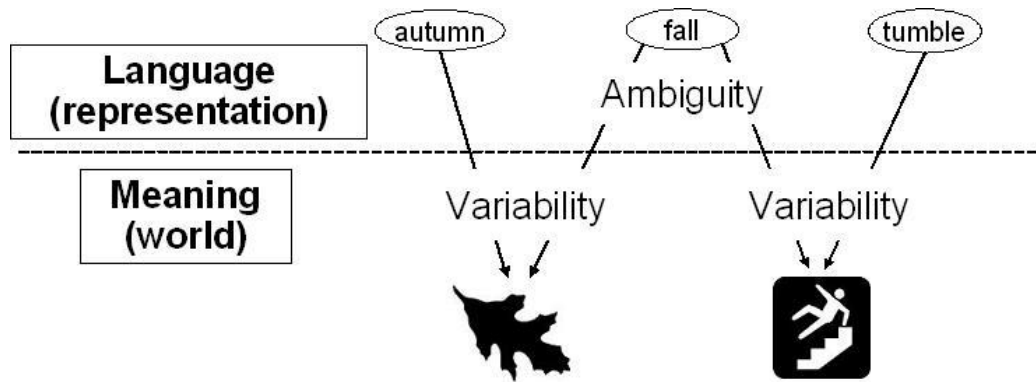


Figure 1.3: Illustration of lexical variability and ambiguity

stipulated senses, logical form, etc.) as a possible mean rather than a goal – we shouldn’t create a-priori artificial problems, which might be harder than those we need to solve.

1.2 Addressing Uncertainty

Uncertainty is yet another issue which needs to be addressed when dealing with inferences in general, and with textual inferences in particular. Once again, uncertainty may arise due to a number of factors. It is most common that one fact may often, but not necessarily, follow another in the world and thus the inference itself is uncertain. Furthermore, given the ambiguity and variability of natural language, it is often the case that the interpretation itself is uncertain. Going back to example C in figure 1.1, there is a linguistically possible reading of the text in which the word ‘some’ in the last sentence refers to ‘retired Canadians’ rather than the more plausible reference to ‘Monarch butterflies’. Another source of uncertainty is due to assumed world knowledge. For example, if one really wants to play devils advocate, one may reason that it does not necessarily follow from the text that the butterflies travel by flying. Given such readings and reasoning, the hypothesis is not absolutely necessarily true but

rather most probably true.

Concrete probabilistic settings made a significant impact on other core NLP tasks such as MT (Brown et al., 1993), IR (Robertson and Jones, 1988; Ponte and Croft, 1998) and syntactic processing (Church, 1988; Collins, 1997). Following this success, this thesis proposes a general generative probabilistic setting that formalizes the notion of probabilistic textual entailment, which enables treating all underlying reasons of uncertainty in uniform manner.

1.3 Thesis Highlights and Contributions

Following are the key and novel contributions of this thesis.

The textual entailment task

The primary contribution of this thesis is in introducing a new application independent empirical task we termed *textual entailment*. Textual entailment abstracts common semantic inferences as needed for text based NLP applications. In Chapter 3 we describe the textual entailment recognition task. Within this application independent framework a text t is said to textually entail a hypothesis h if the truth of h can be most likely inferred from t .

In addition we created an evaluation dataset and made it available to the research community. Up to this point many approaches for modeling textual entailment were developed in application specific settings. The constructed dataset serves as a common benchmark for researchers from different disciplines within NLP to work on a common problem, share techniques and compare results.

The lexical reference subtask

Textual entailment recognition is a complex task that requires deep language understanding. Nevertheless, we decided to focus, as a natural starting point, on relatively shallow inferences and models which operate at the lexical level. We thus define in Chapter 3 also a textual entailment subtask which we term *lexical reference* and devise a corresponding evaluation dataset. We further demonstrate how textual entailment can be approximated by a lexical reference model. Decomposing the textual entailment task into subtasks allows for better analysis as well as better system performance evaluation.

A probabilistic setting for textual entailment

Another salient contribution of this work is the introduction of a general probabilistic setting that formalizes the notion of textual entailment (Chapter 4). We suggest that the proposed setting may provide a unifying framework for modeling uncertain semantic inferences from texts. Presented loosely, we say that a text (probabilistically) textually entails a hypothesis if the text increases the likelihood of the hypothesis being true.

Lexical Models

In Chapter 5 we demonstrate the relevance of the general probabilistic setting for modeling entailment at the lexical level, addressing the lexical reference subtask by devising two derived models that utilize lexical co-occurrence probabilities. The first model assumes an underlying alignment between the text and the hypothesis while the second model views lexical reference as a text classification task. In the second model entailment is derived from the entire context of the sentence (rather than word-to-word alignment) and Naïve Bayes classification is applied in an unsupervised setting to

estimate reference. The proposed probabilistic setting and the derived initial models demonstrate a promising direction for additional improved probabilistic models for textual entailment.

Acquisition of lexical entailment relations

Chapter 6 makes a slight shift from inference to acquisition and presents a novel algorithm for learning lexical entailment relations focusing on the extraction of verb paraphrases. Most previous approaches detect individual paraphrase instances within comparable corpora, each of them containing roughly the same information, and rely on the substantial level of correspondence of such corpora. Our proposed method successfully detects isolated paraphrase instances within a single corpus without relying on any a-priori structure or information. We propose a novel instance based approach which identifies actual paraphrase instances that describe the same fact or event rather than comparing typical contexts in a global manner as in vector-based similarity approaches. Our instance based approach seems to address some of the drawbacks of distributional similarity based equivalents, in particular by providing a consistent scoring scale across different words.

“The more extensive a man’s knowledge of what has been done, the greater will be his power of knowing what to do.”

— Benjamin Disraeli

Chapter 2

Background

2.1 Application Needs

Inferencing between language expressions is at the heart of many high level natural language processing tasks including Question Answering, Information Retrieval and Extraction and others that attempt to reason about and capture the meaning of linguistic expressions. In the following subsections we describe how semantic inferencing and modeling of textual variability are related to specific core NLP applications.

2.1.1 Question Answering (QA)

Aiming at returning brief answers in response to natural language questions, open-domain QA systems represent an advanced application of natural language processing (Hirschman and Gaizauskas, 2001). The typical setting of QA addresses the task of finding answers to natural language questions (e.g. “How tall is the Eiffel Tower?” “Who is Aaron Copland?”) from large text collections. Reading comprehension tests

provide an additional setting for question answering, based on a system’s ability to answer questions about a specific reading passage (Hirschman et al., 1999).

Finding answers requires processing texts at a level of detail that cannot be carried out at retrieval time for very large text collections. Due to this limitation most QA systems apply a preceding retrieval phase in which a subset of query-relevant texts are selected from the whole collection¹. However, a question might be formulated using certain words and expressions while answer texts, to be found in a corpus, might include variations of the same expressions. As it turns out, modeling semantic variability plays a much more important role in QA retrieval than in standard topical IR – in QA the query is usually longer and the chance of finding text segments that closely match all question tokens is small.

Given a set of candidate answer texts, a typical QA system must apply reasoning to identify which texts are informative answer texts. For example, given the question “*What is the height of the Eiffel Tower?*” textual inferencing must be done to infer that text such as “*The Eiffel tower is 300 meters tall*” constitute an informative answer to the question while text such as “*The Eiffel tower is 115 years old*” do not.

Another interesting role for modeling semantic inference in QA is answer fusion in which semantic relations between answer texts can be exploited for better presentation of results (Dalmas and Webber, 2005). For example, given the question “*Where is the Taj Mahal?*” it is useful to identify that answer texts regarding Agra or India are related and are distinct from texts on the Trump Taj Mahal Casino Resort in Atlantic City.

2.1.2 Information Extraction (IE)

IE can be defined as the task of filling predefined templates from natural language texts, where the templates are designed to capture information about key role players

¹see dedicated workshop on the IR4QA subtask: <http://nlp.shef.ac.uk/ir4qa04/>

in stereotypical events (Gaizauskas and Wilks, 1998). For example, a template can be defined to capture information about corporate takeover events; such a template has slots for the acquiring company, the acquired company, the the amount paid, etc. Once again, an IE system needs to identify the various ways in which such a relation could be expressed. Contrary to QA and ad-hoc IR, the static templates are given and thus the IE task is commonly addressed with supervised learning techniques.

2.1.3 Information Retrieval (IR)

The primary task of information retrieval is to retrieve a set of documents that are *relevant* for an information need specified by a search query, typically consisting one or more natural language terms. Most users of document retrieval systems when formulating their query usually employ terms that they expect to appear in a relevant document. However, a document may not contain all query terms and yet be relevant. For example, a document about ‘*unix*’ may be relevant to a query about “*operating systems*,” yet the words ‘operating’ or ‘system’ may be absent in that document. An IR system would definitely want to address this scenario and identify when a document is relevant regardless of the occurrence or absence of the query tokens in the document. Modeling language variability is thus an important factor in IR.

This lexical gap between the query and document is typically addressed by lexical expansion techniques (see Section 2.2.1). However, some works have addressed the foundational problem of IR by trying to give a formal notion of relevance based on mathematical logic. Van Rijsbergen (1979) views the IR task as inferring the query based on the text. According to this proposal, in order to estimate the relevance of a document D with respect to a query Q , we have to estimate the probability that D implies Q . IR is thus viewed as essentially consisting of a disguised form of logical inference.

2.1.4 Summarization

Radev (2000) describes 24 cross-document relations that can hold between segments of documents, one of which is the entailment relation. It can be used to compute the informativity of one text segment compared to another one. In the context of summarization this is used to avoid redundancy, i.e., if a segment entails another segment, only the entailing segment should be included in the summary. In particular, Multi-document summarization systems need to deduce that different expressions found in several documents express the same meaning; hence only one of them should be included in the final summary (Barzilay, McKeown, and Elhadad, 1999).

2.1.5 Generation

Natural Language Generation (NLG) commonly refers to the task of generating natural language from a machine representation system such as a knowledge base or a logical form. However in many application settings there is a need for reformulating or paraphrasing natural language texts. For example, the *sentence compression* task involves simplifying sentences by removing lexical material or by paraphrasing parts of them (Knight and Marcu, 2002). For example, due to time and space constraints, the generation of TV captions often requires only the most important parts of sentences to be shown on a screen (Daelemans, Höthker, and Tjong Kim Sang, 2004). Sentence simplification involves capturing the original text's most salient pieces of information and reformulating it with minimum loss of information.

2.1.6 Automatic Machine Translation (MT) Evaluation

MT evaluation faces the problem that there is no single good translation. Typically, there are many acceptable translations of a given source sentence. These translations may vary in word choice or in word order even when they use the same words. Human

evaluation of system output is costly in both time and money, leading to the rise of automatic evaluation metrics in recent years (Papineni et al., 2001). The automatic MT evaluation task involves comparing the output of an MT system and one or more reference translations. The output of an MT system should capture the meaning of the original text as denoted in the reference translations.

2.2 Techniques and Methods

Within application settings a wide variety of techniques were proposed to address semantic variability, ranging at different levels of representation and complexity.

2.2.1 Thesaurus-based Term Expansion

Thesaurus-based term expansion (or substitution) is a commonly used technique for enhancing the recall of NLP systems and coping with lexical variability. Expansion consists of altering a given text (usually a query) by adding terms of similar meaning. WordNet (Miller, 1995) is commonly used as a source of related words for expansion.

For example, many QA systems perform expansion in the retrieval phase using query related words based on WordNet’s lexical relations. In some QA systems, (Harabagiu et al., 2000; Kwok, Etzioni, and Weld, 2001; Hovy, Hermjakob, and Lin, 2001), there is no clear weighting scheme for the lexical expansion, and expanded words are added equally to a Boolean retrieval search of candidate answer texts. Sag-gion et al. (2004) do propose ranking the candidate answer texts for a given question based on the degree of word overlap between the question q and the candidate text t . The proposed measure takes into account the the inverse document frequency (*idf*) of the words as follows: $score(q, t) = \sum_{u \in q \wedge t} idf(u)$. This measure favors texts that were retrieved by fewer expansions or by expansions of relatively frequent words. In (Yang and Chua, 2002; Negri, 2004; Matthew W. Bilotti and Lin, 2004), WordNet

expansion is integrated into the retrieval model, in which the expanded words weights are discounted.

Within the context of IE , Califf and Mooney (2003) successfully integrate WordNet expansions into the supervised learning mechanism. Chai and Bierman (1997) utilize WordNet information to help users customize an extraction system to satisfy their particular needs.

Query expansion with WordNet is believed to be potentially relevant to enhance recall in IR , as it permits matching relevant documents that do not contain any of the query terms (Voorhees, 1994). However, despite some positive results (e.g. (Nie and Brisebois, 1996)) it has produced mostly unsuccessful experiments.

2.2.2 Distributional Similarity

Distributional Similarity between words has been an active research area for more than a decade. It is based on the general idea of Harris' Distributional Hypothesis, suggesting that words that occur within similar contexts are semantically similar (Harris, 1968). Concrete similarity measures compare a pair of weighted context feature vectors that characterize two words (e.g. (Lin, 1998)).

Qiu and Frei (1993) present a probabilistic query expansion model based on a similarity thesaurus which was constructed automatically and show that this kind of query expansion results in a notable improvement in the retrieval effectiveness.

However, distributional similarity does not capture equivalence and entailment of meaning but rather broader meaning similarity. For this reason distributional similarity is not commonly used within applications.

Geffet and Dagan (2004) and Geffet and Dagan (2005) try to address this issue by proposing a new feature weighting function and a new vector comparison scheme that yield more accurate distributional similarity lists, which better approximate equivalence and entailment of meaning.

2.2.3 Lexical Overlap

Modeling lexical overlap between texts is a common problem in NLP applications. Many techniques and heuristics were applied within various applications to model such relations between text segments. Within the context of Multi Document Summarization, Monz and de Rijke (2001) propose modeling the directional entailment between two texts t , h to identify redundant information appearing in different texts via the following entailment score:

$$entscore(t, h) = \frac{\sum_{w \in t \wedge h} idf(w)}{\sum_{w \in h} idf(w)} \quad (2.1)$$

where $idf(w) = \log(N/n_w)$, N is the total number of documents in the corpus and n_w the number of documents containing word w . A practically equivalent measure was independently proposed in the context of QA to rank query expansion retrieval results (see Section 2.2.1). This baseline measure captures word overlap, considering only words that appear in both texts and weighs them based on their inverse document frequency.

The BLEU (BiLingual Evaluation Understudy) algorithm (Papineni et al., 2001) is a method for automatic evaluation of machine translation. Basically, the algorithm looks for n-gram coincidences between a candidate text (the automatically produced translation) and a set of reference texts (the human-made translations). ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a slightly modified version of the BLEU algorithm which has been applied to evaluate text summarization systems (Lin, 2004).

2.2.4 Paraphrase Acquisition

Recently, several works addressed the task of acquiring paraphrases (semi-) automatically from corpora. Most attempts were based on identifying corresponding sentences

in parallel or ‘comparable’ corpora, where each corpus is known to include texts that largely correspond to texts in another corpus (see next section). The major types of comparable corpora are different translations of the same text, and multiple news sources that overlap largely in the stories that they cover. Typically, such methods first identify pairs (or sets) of larger contexts that correspond to each other, such as corresponding documents, by using clustering or similarity measures at the document level, and by utilizing external information such as requiring that corresponding documents will be from the same date. Then, within the corresponding contexts, the algorithm detects individual pairs (or sets) of sentences that largely overlap in their content and are thus assumed to describe the same fact or event. Barzilay and McKeown (2001) use sentence alignment to identify paraphrases from a corpus of multiple English translations of the same text. Pang, Knight, and Marcu (2003) also use a parallel corpus of Chinese-English translations to build finite state automata for paraphrase patterns, based on syntactic alignment of corresponding sentences. Shinyama et al. (2002) learn structural paraphrase templates for information extraction from a comparable corpus of news articles from different news sources over a common period of time. Similar news article pairs from the different news sources are identified based on document similarity. Sentence pairs are then identified based on the overlap of named entities in the matching sentences. Barzilay and Lee (2003) also utilizes a comparable corpus of news articles to learn paraphrase patterns, which are represented by word lattice pairs. Patterns originating from the same day but from different newswire agencies are matched based on entity overlap.

Lin and Pantel (2001) propose a different approach for extracting ‘inference rules’, which largely correspond to paraphrase patterns. Their method extracts such paraphrases from a single corpus rather than from a comparable set of corpora. It is based on vector-based similarity (Lin, 1998), which compares typical contexts in a global manner rather than identifying all actual paraphrase instances that describe the same

fact or event. The underlying assumption in their work is that paths in dependency trees that connect similar syntactic arguments (slots) are close in meaning. Rather than considering a single feature vector that originates from the arguments in both slots, at both ends of the path, vector-based similarity was computed separately for each slot. The similarity of a pair of binary paths was defined as the geometric mean of the similarity values that were computed for each of the two slots.

TEASE (Szpektor et al., 2004) is a scalable unsupervised web-based method for extracting candidate entailment relations. It requires as input a lexicon of core terms for which paraphrases are sought, without any additional supervision. The system then extracts reliable anchor sets for a given core term from sample texts retrieved from the web. An anchor set is a set of context terms which indicates with a high probability that a common fact is described in multiple sentences. Iterative web search queries are performed to retrieve first sentences containing the core term, and then to retrieve sentences containing the associated anchors. Statistical criteria are applied over the retrieved anchor candidates to identify promising anchor sets. As a second phase the system applies a symbolic structure learning algorithm over the sentences which contain the anchor sets. This algorithm identifies the most general (smallest) linguistic structures, called templates, across multiple anchor sets that connect anchors in the parsed sentences, resembling ILP-style symbolic learning. The output templates are obtained by replacing the anchors with variables in the structures, and are then assumed to be entailing and/or entailed by the core lexical relation. For example, for the verb *prevent*, the system correctly identifies corresponding templates such as “reduce the risk of.”

A related probabilistic approach for paraphrasing (Quirk, Brockett, and Dolan, 2004) apply the classical statistical machine translation model, implemented by the GIZA++ software (Och and Ney, 2003), in which probabilistic “paraphrase” rules are captured mostly by translation probabilities for lexical elements (and some additional

parameters of the statistical translation model).

2.2.5 Mapping to Logical Form

Though the bulk of state of the art work consists of model and processing text at a lexical or shallow syntactic level, throughout the years there were many approaches to apply deep semantic inferences based on mapping text to some intermediate level of logical representation. For example, Blackburn and Bos (2005) propose and describe how to interpret language expressions to a logical interpretation and to perform inference with the result in terms of theorem proving.

Within the context of QA, Crouch et al. (2003) utilize deep semantic representations by interpreting text into a formal language on which inference is performed. Moldovan and Rus (2001) present a method for transforming the WordNet glosses into logic forms and further into axioms used for theorem proving. The paper demonstrates the utility of the axioms in a question answering system to rank and extract answers. For example, to prove that “. . . Socrates’ death came when he chose to drink poisoned wine. . .” is a plausible answer to the question “How did socrates die?”, one needs to know that drinking poisoned wine may be a cause of death. This extra knowledge is found in WordNet in the gloss of the second sense of poison (“kill with poison”) and in the first sense of kill (“cause to die”) which collectively justify the answer.

Hobbs et al. (1988) have presented models of the interpretation process based on weighted abduction. In the process of text comprehension, utterances are understood to provide only partial information as to utterance meaning, and some information must be abduced.

While addressing inference tasks in computational semantics by means of first-order theorem proving tools is an important and welcome development, it has some

inherent limitations. First, generating first-order logic representations of natural language documents is hampered by the lack of efficient and sufficiently robust NLP tools. Second, the computational costs of deploying first-order logic theorem proving tools in real-world situations may be prohibitive. And third, the strict yes/no decisions delivered by such tools are not always appropriate.

2.3 Summary

In summary, many approaches for modeling semantic variability and entailment were developed in application specific settings. Overall, the common practice is mostly shallow processing, based on measuring lexical overlap or modeling lexical variability (usually from a thesaurus) or by applying simple reformulation patterns. Furthermore, in an abstract application-independent setting it is not clear how scores for semantic variations should be assigned and interpreted, which may call for a common evaluation dataset and a generic probabilistic setting for textual entailment.

on the Arthur Bernstein piano competition

“... Competition, even a piano competition, is legitimate ... as long as it is just an anecdotal side effect of the musical culture scene, and doesn’t threat to overtake the center stage”

— Haaretz News Paper, April 2005

Chapter 3

The Applied Textual Entailment Recognition Task

Even though different applications need similar models for semantic variability, the problem is often addressed in an application-oriented manner and methods are evaluated by their impact on final application performance (see Chapter 2). Consequently it becomes difficult to compare, under a generic evaluation framework, practical inference methods that were developed within different applications. Furthermore, researchers within one application area might not be aware of relevant methods that were developed in the context of another application. Overall, there seems to be a lack of a clear framework of generic task definitions and evaluations for such “applied” semantic inference, which also hampers the formation of a coherent community that addresses these problems. This situation might be confronted, for example, with the state of affairs in syntactic processing, where clear application-independent tasks,

communities (and even standard conference session names) have matured.

3.1 Task Definition

Recently there have been just a few suggestions in the literature to regard entailment recognition for texts as an applied, empirically evaluated, task (see (Crouch et al., 2003; Dagan and Glickman, 2004; Monz and de Rijke, 2001)). The *textual entailment* task is an attempt to promote an abstract generic task that captures major semantic inference needs across applications. The task requires to recognize, given two text fragments, whether the meaning of one text can be inferred (entailed) from another text. More concretely, our applied notion of textual entailment is defined as a directional relationship between pairs of text expressions, an entailing *text*, and an entailed textual *hypothesis* as follows:

Definition 1 *We say that a text T entails a hypothesis H if, typically, a human reading T would infer that H is most likely true.*

This somewhat informal definition (a concrete version of detailed guidelines follows) is based on and assumes common human understanding of language as well as common background knowledge. It is similar in spirit to evaluation of applied tasks such as QA and IE, in which humans need to judge whether the target answer or relation can indeed be inferred from a given candidate text. As in other evaluation tasks our definition of textual entailment is operational, and corresponds to the judgment criteria given to the annotators who decide whether this relationship holds between a given pair of texts or not. We explicitly include “fuzzy” terms such as *typically* and *most likely* as this seems to characterize empirical evaluation criteria and task definition (e.g. (Voorhees and Tice, 2000)) and is necessary given the uncertain nature of our task.

Our applied notion of textual entailment is related, of course, to classical semantic entailment in the linguistics literature. A common definition of entailment in formal semantics (Chierchia and McConnell-Ginet, 2000) specifies that a text t entails another text h (hypothesis, in our terminology) if h is true in every circumstance (*possible world*) in which t is true. For example, in example 13 from Table 3.1 we'd assume humans to agree that the hypothesis is necessarily true in any circumstance for which the text is true. In such intuitive cases, our proposed notion of textual entailment corresponds to the classical notions of semantic entailment.

However, our applied definition allows for cases in which the truth of the hypothesis is highly plausible, for most practical purposes, rather than certain. In Table 3.1, examples 1586, 1076, 893 and 586 were judged as true by annotators even though the entailment in this cases is not certain. This seems to match the types of uncertain inferences that are typically expected from text based applications.

3.1.1 Judgement Guidelines

An empirical task is equivalent to its annotation guidelines. Following are the concrete textual entailment annotation guidelines for text hypothesis pairs as presented to the annotators.

Read the text. And decide if based on the text the hypothesis is most probably true taking into account the following guidelines:

- Entailment is a directional relation. The hypothesis must be entailed from the given text, but the text need not be entailed from the hypothesis.
- In principle, the hypothesis must be fully entailed by the text. Judgment should be false if the hypothesis includes parts that cannot be inferred from the text.

- Cases in which inference is very probable (but not absolutely certain) should be judged at true. In example #586 in Table 3.1 one could claim that the shooting took place in 1993 and that (theoretically) the cardinal could have been just severely wounded in the shooting and has consequently died a few months later in 1994. However, this example should be tagged as true since the context seems to imply that he actually died in 1993.
- You are allowed to assume common background knowledge, which is typical for a reader of the given type of texts, such as that a company has a CEO, a CEO is an employee of the company, an employee is a person, etc. However, it was considered unacceptable to presume highly specific knowledge, such as that Yahoo bought Overture for 1.63 billion dollars.

3.1.2 Mapping to Applications

It seems that major inferences, as needed by multiple applications, can indeed be cast in terms of our definition of textual entailment (see also Section 2.1).

Question-answering can often be reduced to a textual entailment problem by rephrasing the question as a declarative statement that is to be entailed by correct answers. Whether a candidate text actually answers a question can be inferred by deciding if the converted question is entailed by the candidate answer. For example, given the question “*Who painted ‘The Scream’?*”, the text “*Norway’s most famous painting, ‘The Scream’ by Edvard Munch, . . .*” is an informative answer text since it entails the hypothesized answer form “Edvard Munch painted ‘The Scream’.” (see corresponding $T - H$ pair in example 568 from Table 3.1).

Similarly, for certain IR queries the combination of semantic concepts and relations denoted by the query should be entailed from relevant retrieved documents. For example, given the query “ineffectiveness of antibiotics”, a typical user would

be interested in retrieving documents which entail the meaning of the query (possibly reformulated as the sentential statement “antibiotics is ineffective”). A possible entailing (and relevant) text being: “*Since the common cold is caused by a virus, antibiotics will not cure it.*”

In IE, entailment holds between different text variants that express the same target relation. This can be reformulated as a textual entailment problem by simply phrasing the template as a language expression. For example, to fill an acquisition template, a text should entail “X acquired Y for Z” where company names X and Y fill the acquiring company and acquired company slots and a monetary value Z fills the the amount paid slot.

In multi-document summarization a redundant sentence, to be omitted from the summary, should be entailed from other sentences in the summary. And in MT evaluation a correct translation should be semantically equivalent to the gold standard translation, and thus both translations should entail each other.

Additional examples for recasting applications in terms of textual entailment are described in Section 3.2.2. Consequently, we suggest that textual entailment recognition is a suitable generic task for evaluating and comparing applied semantic inference models. Eventually, such efforts can promote the development of entailment recognition “engines” which may provide useful generic modules across applications.

3.2 Creating the PASCAL Evaluation Dataset

We created a textual entailment public benchmark within the framework of the PASCAL challenges (Dagan, Glickman, and Magnini, 2005). The dataset consists of text-hypothesis ($T-H$) pairs of small text snippets, corresponding to the general news domain. Examples were manually labeled for entailment - whether T entails H or not - by human annotators, and were divided into *development* and *test* datasets.

ID	TEXT	HYPOTHESIS	TASK	VALUE
568	<i>Norway’s most famous painting, “The Scream” by Edvard Munch, was recovered Saturday, almost three months after it was stolen from an Oslo museum.</i>	<i>Edvard Munch painted “The Scream”.</i>	QA	True
1586	<i>The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.</i>	<i>The national language of Yemen is Arabic.</i>	QA	True
1076	<i>Most Americans are familiar with the Food Guide Pyramid– but a lot of people don’t understand how to use it and the government claims that the proof is that two out of three Americans are fat.</i>	<i>Two out of three Americans are fat.</i>	RC	True
1667	<i>Regan attended a ceremony in Washington to commemorate the landings in Normandy.</i>	<i>Washington is located in Normandy.</i>	IE	False
13	<i>iTunes software has seen strong sales in Europe.</i>	<i>Strong sales for iTunes in Europe.</i>	IR	True
2016	<i>Google files for its long awaited IPO.</i>	<i>Google goes public.</i>	IR	True
2097	<i>The economy created 228,000 new jobs after a disappointing 112,000 in June.</i>	<i>The economy created 228,000 jobs after disappointing the 112,000 of June.</i>	MT	False
893	<i>The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD.</i>	<i>The first settlements on the site of Jakarta were established as early as the 5th century AD.</i>	CD	True
1960	<i>Bush returned to the White House late Saturday while his running mate was off campaigning in the West.</i>	<i>Bush left the White House.</i>	PP	False
586	<i>The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.</i>	<i>Cardinal Juan Jesus Posadas Ocampo died in 1993.</i>	QA	True
908	<i>Time Warner is the world’s largest media and Internet company.</i>	<i>Time Warner is the world’s largest company.</i>	RC	False
1911	<i>The SPD got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%.</i>	<i>The SPD is defeated by the opposition parties.</i>	IE	True

Table 3.1: Examples of text-hypothesis pairs from the PASCAL Recognising Textual Entailment Challenge Dataset

Table 3.1 includes a few examples from the dataset along with their gold standard annotation.

The dataset was collected with respect to different text processing applications, as detailed in the next section. Each portion of the dataset was intended to include typical T - H examples that may correspond to success and failure cases of the actual applications. The collected examples represent a range of different levels of entailment reasoning, based on lexical, syntactic, logical and world knowledge, at different levels of difficulty.

The distribution of examples in this dataset has been somewhat biased to choosing nontrivial pairs, and also imposed a balance of true and false examples. For this reason, systems performances in applicative settings might be different than the figures for this dataset, due to different distributions of examples in particular applications. Yet, the dataset does challenge systems to handle properly a broad range of entailment phenomena. Overall, we were hoping that meaningful baselines and analyses for the capabilities of current systems will be obtained.

3.2.1 Dataset Preparation

The dataset of text-hypothesis pairs was collected by human annotators. It consists of seven subsets, which correspond to typical success and failure settings in different applications, as listed below. Within each application setting the annotators selected both positive entailment examples (*true*), where T is judged to entail H , as well as negative examples (*false*), where entailment does not hold (a 50%-50% split). Typically, T consists of one sentence (sometimes two) while H was often made a shorter sentence (see Table 3.1). The full datasets are available online¹.

In some cases the examples were collected using external sources, such as available datasets or systems (see Acknowledgements), while in other cases examples were

¹<http://www.pascal-network.org/Challenges/RTE/>

collected from the web, focusing on the general news domain. In all cases the decision as to which example pairs to include was made by the annotators. The annotators were guided to obtain a reasonable balance of different types of entailment phenomena and of levels of difficulty. Since many T - H pairs tend to be quite difficult to recognize, the annotators were biased to limit the proportion of difficult cases, but on the other hand to try avoiding high correlation between entailment and simple word overlap. Thus, the examples do represent a useful broad range of naturally occurring entailment factors. Yet, we cannot say that they correspond to a particular representative distribution of these factors, or of true vs. false cases, whatever such distributions might be in different settings. Thus, results on this dataset may provide useful indications of system capabilities to address various aspects of entailment, but do not predict directly the performance figures within a particular application.

As a general guideline in the preparation annotators were guided to avoid vague examples for which inference has some positive probability that is not clearly very high. In addition, to keep the contexts in T and H self-contained annotators were allowed to replace anaphors with the appropriate reference from preceding sentences where applicable. They were also guided to shorten the hypotheses, and sometimes the texts, to reduce complexity.

It is interesting to note in retrospect that the annotators' selection policy yielded more negative examples than positive ones in the cases where T and H have a very high degree of lexical overlap. This anomaly was noticed also by (Bos and Markert, 2005; Bayer et al., 2005) and is further discussed in Chapter 5.

3.2.2 Application Settings

Following we describe the specific preparation procedures and guidelines for different application settings. Text hypothesis examples for the various tasks can be found in Table 3.1 (see *task* column).

Comparable Documents (CD)

Annotators identified T - H pairs by examining a cluster of comparable news articles that cover a common story. They examined “aligned” sentence pairs that overlap lexically, in which semantic entailment may or may not hold. Some pairs were identified on the web using Google news¹ and others taken a corpus of sentence alignment in monolingual comparable corpora from Columbia University². The motivation for this setting is the common use of lexical overlap as a hint for semantic overlap in comparable documents, e.g. for multi-document summarization.

Reading Comprehension (RC)

This task corresponds to a typical reading comprehension exercise in human language teaching, where students are asked to judge whether a particular assertion can be inferred from a given text story. The annotators were asked to create such hypotheses relative to texts taken from news stories, considering a reading comprehension test for high school students.

Question Answering (QA)

Annotators used the TextMap web-based question answering system available online³. The annotators used questions from CLEF-QA⁴ (mostly) and TREC⁵, but could also construct their own questions. For a given question, the annotators chose first a relevant text snippet (T) that was suggested by the QA system as including the correct answer. They then turned the question into an affirmative sentence with the hypothesized answer “plugged in” to form the hypothesis (H). For example, given

¹<http://news.google.com>

²<http://www.cs.columbia.edu/~noemie/alignment/>

³<http://brahms.isi.edu:8080/textmap/>

⁴<http://clef-qa.itc.it/>

⁵<http://trec.nist.gov/data/qa.html>

the question, “Who is Ariel Sharon?” and taking a candidate answer text “*Israel’s Prime Minister, Ariel Sharon, visited Prague*” (T), the hypothesis H is formed by turning the question into the statement “*Ariel Sharon is Israel’s Prime Minister*”, producing a true entailment pair.

Information Extraction (IE)

This task is inspired by the Information Extraction application, adapting the setting for pairs of texts rather than a text and a structured template. For this task the annotators used an available dataset annotated for the IE relations “kill” and “birth place” produced in the University of Illinois at Urbana-Champaign¹, as well as general news stories in which the annotators identified manually typical IE relations. Given an IE relation of interest (e.g. a purchasing event), annotators identified as the text (T) candidate news story sentences in which the relation is suspected to hold. As a hypothesis they created a straight-forward natural language formulation of the IE relation, which expresses the target relation with the particular slot variable instantiations found in the text. For example, given the information extraction task of identifying killings of civilians, and a text “*Guerrillas killed a peasant in the city of Flores.*”, a hypothesis “*Guerrillas killed a civilian*” is created, producing a true entailment pair.

Machine Translation (MT)

We used the Document Understanding Conference (DUC) 2004 machine translation evaluation data, from the National Institute of Standards and Technology (NIST)². Two translations of the same text, an automatic translation and a gold standard human translation were compared and possibly modified in order to obtain T - H pairs.

¹<http://12r.cs.uiuc.edu/~cogcomp/>

²<http://duc.nist.gov/duc2004/>

The automatic and gold standard translation were alternately taken as either T or H , where a correct translation corresponds to true entailment. The automatic translations were sometimes grammatically adjusted when being otherwise grammatically unacceptable.

Information Retrieval (IR)

Annotators generated hypotheses (H) that may correspond to meaningful IR queries that express some concrete semantic relations. These queries are typically longer and more specific than a standard keyword query, and may be considered as representing a semantic-oriented variant within IR. The queries were selected by examining prominent sentences in news stories, and were then submitted to a web search engine. Candidate texts (T) were selected from a web search engine's retrieved documents, picking candidate text snippets that either do or do not entail the hypothesis.

Paraphrase Acquisition (PP)

Paraphrase acquisition systems attempt to acquire pairs (or sets) of lexical-syntactic expressions that convey largely equivalent or entailing meanings. This task is related to generation systems in which sentences need to be paraphrased.

We used the following two sources of paraphrase databases. The output of the TEASE system for extracting entailment relations and paraphrases (Szpektor et al., 2004) and the DIRT paraphrase database¹. Annotators selected a text T from some news story which includes a certain lexical-syntactic relation, for which a paraphrase rule from the paraphrase acquisition system may apply. The result of applying the paraphrase rule on T was chosen as the hypothesis H . Correct paraphrases suggested by the system, which were applied in an appropriate context, yielded true

¹<http://www.isi.edu/~pantel/Content/Demos/demosDirt.htm>

T - H pairs; otherwise a false example was generated. For example, given the sentence “*The girl was found in Drummondville.*” and by applying the paraphrase rule X was found in $Y \Rightarrow Y$ contains X , we obtain the hypothesis “*Drummondville contains the girl.*” yielding a false example.

3.2.3 The Annotation Process

Each example T - H pair was first judged as true/false by the annotator that created the example. The examples were then cross-evaluated by a second judge, who received only the text and hypothesis pair, without any additional information from the original context. Judgements were based on the guidelines of Section 3.1.1. However, given that the text and hypothesis might originate from documents at different points in time, annotators were asked to ignore tense aspects.

The annotators agreed in their judgment for roughly 80% of the examples, which corresponded to a Kappa of 0.6 which is regarded as moderate agreement (Landis and Koch, 1997)). The 20% of the pairs for which there was disagreement among the judges were discarded from the dataset. Furthermore, the author performed a light review of the remaining examples and eliminated an additional 13% of the original examples, which might have seemed controversial. Altogether, about 33% of the originally created examples were filtered out in this process.

The remaining examples were considered as the gold standard for evaluation, split to 567 examples in the development set and 800 in the test set, and evenly split to true/false examples. Our conservative selection policy aimed to create a dataset with non-controversial judgments, which will be addressed consensually by different groups. It is interesting to note that portions of the dataset have independently been judged by other groups who reached high agreement levels with the gold standard judgments, of 95% on all the test set (Bos and Markert, 2005), 96% on a subset of roughly a third of the test set (Vanderwende, Coughlin, and Dolan, 2005) and 91%

on a sample of roughly 1/8 of the development set (Bayer et al., 2005).

3.3 Evaluation Measures

Given that in our dataset we used a binary {true, false} annotation, textual entailment recognition can be seen as a binary classification task. Judgments (classifications) produced by a systems can be compared to the gold standard and the percentage of matching judgments provides the accuracy of the run, i.e. the fraction of correct responses. Accuracy is a standard measure in the machine learning community and is commonly used to evaluate performance in many NLP tasks such as text classification, information extraction and word sense disambiguation. In our setting, since the dataset was balanced in terms of true and false examples, a system that uniformly predicts true (or false) would achieve an accuracy of 50% which constitutes a natural baseline. Another baseline is obtained by considering the distribution of results in random runs that predict true or false at random. A run on the test set with *accuracy* > 0.535 is better than chance at the 0.05 significance level and a run with *accuracy* > 0.546 is better than chance at the 0.01 level¹.

In many learning problems in general and NLP tasks in particular, the goal is not simply to classify objects into one of a fixed number of classes; instead, a ranking of objects is desired. This is the case in information seeking applications such as IR or QA where one is interested in retrieving texts from some collection that are relevant to a given query or question. In such problems, one wants to return a list of texts that contains the relevant texts at the top and irrelevant texts at the bottom; in other words, one wants a ranking of the texts such that relevant ones are ranked higher than irrelevant ones. For such precision oriented tasks, the recall-precision graph is quite informative. *Average Precision* is a single figure measure equivalent to to the area

¹These figures were obtained by simulating ten thousand random runs

under an uninterpolated recall-precision curve and is commonly used to evaluate a systems ranking ability (Voorhees and Harman, 1999). Average precision emphasizes ranking true examples before false ones. As a second measure, we propose average precision for evaluating textual entailment recognition. The measure is defined, in our case, as the sum of the precision at each true example in the ranked data divided by the total number of true examples in the collection as follows:

$$\begin{aligned} \text{average precision} &= \frac{\sum_{i=1}^N P(i)T(i)}{\sum_{i=1}^N T(i)} \\ P(i) &= \frac{\sum_{k=1}^i T(k)}{i} \end{aligned} \tag{3.1}$$

where N is the number of the pairs in the test set (800 in our case), $T(i)$ is the gold annotation (true=1, false=0) and i ranges over the ranked pairs. An average precision of 1.0 means that the system assigned a higher score to all true examples than to any false one (perfect ranking). For the our test test the worst possible system, which ranks all false examples before any true one, will achieve an average precision of 0.3.

We note that the machine learning community recognizes as well the needs of precision oriented tasks for evaluation measures other than accuracy and offer learning methods that can directly optimize performance measures like average precision (e.g. (Joachims, 2005)).

We would like to emphasize that a test set of 800 examples, while being large enough for reporting statistically significant results, is definitely not large enough to capture the whole wealth of lexical choice, syntactic constructs and entailment phenomena. Reported results on this dataset should be perceived appropriately. For a better understanding of system performance additional datasets for the overall task and for specific subtasks should be considered.

3.4 Dataset Analysis

The resulting dataset was the basis for the PASCAL Network of Excellence Recognising Textual Entailment Recognition (RTE-1) Challenge. In this section we present insights and analysis of the dataset which were obtained based on the performance of systems participating in the challenge. Submitted systems were asked to tag each $T-H$ pair as either true, predicting that entailment does hold for the pair, or as false otherwise. In addition, systems could optionally add a confidence score (between 0 and 1) where 0 means that the system has no confidence of the correctness of its judgment, and 1 corresponds to maximal confidence. Systems were evaluated based on accuracy and *confidence weighted score* (Dagan, Glickman, and Magnini, 2005). Note that in the calculation of the confidence weighted score¹ correctness is with respect to classification - i.e. a negative example, in which entailment does not hold, can be correctly classified as false. This is slightly different from our proposed use of the average precision measure (as common in IR and QA), in which systems rank the results by confidence of positive classification and correspondingly only true positives are considered correct. The evaluation scheme proposed here has been adopted and is planned for the forthcoming RTE-2 challenge.

The development data set was intended for any system tuning needed. It was acceptable to run automatic knowledge acquisition methods (such as synonym collection) specifically for the lexical and syntactic constructs present in the test set, as long as the methodology and procedures are general and not tuned specifically for the test data.

Sixteen groups submitted the results of their systems for the challenge data, while one additional group submitted the results of a manual analysis of the dataset (Vanderwende, Coughlin, and Dolan, 2005). The submitted systems incorporated a broad

¹ $cws = \frac{1}{n} \sum_{i=1}^n \frac{\#correct-up-to-rank-i}{i}$

First Author (Group)	accuracy	cws	partial coverage	System description					
				Word overlap	Statistical lexical relations	WordNet	Syntactic matching	world knowledge	Logical inference
Akhmatova (Macquarie)	0.519	0.507		X					X
Andreevskaja (Concordia)	0.519	0.515				X	X		
	0.516	0.52							
Bayer (MITRE)	0.586	0.617			X				
	0.516	0.503	73%					X	X
Bos (Edinburgh & Leeds)	0.563	0.593		X		X		X	X
	0.555	0.586		X					
Delmonte (Venice & irst)	0.606	0.664	62%			X	X		X
Fowler (LCC)	0.551	0.56				X		X	X
Glickman (Bar Ilan)	0.586	0.572			X				
	0.53	0.535							
Herrera (UNED)	0.566	0.575		X	X		X		
	0.558	0.571		X					
Jijkoun (Amsterdam)	0.552	0.559		X	X				
	0.536	0.553		X		X			
Kouylekov (irst)	0.559	0.607		X	X		X		
	0.559	0.585							
Newman (Dublin)	0.563	0.592		X	X				
	0.565	0.6							
Perez (Madrid)	0.495	0.517		X					
	0.7	0.782	19%						
Punyakank (UIUC)	0.561	0.569					X		
Raina (Stanford)	0.563	0.621			X	X	X		X
	0.552	0.686							
Wu (HKUST)	0.512	0.55			X		X		
	0.505	0.536							
Zanzotto (Rome-Milan)	0.524	0.557				X	X		
	0.518	0.559							

Table 3.2: Accuracy and cws results for the system submissions, ordered by first author. Partial coverage refers to the percentage of examples classified by the system out of the 800 test examples.

range of inferences that address various levels of textual entailment phenomena. Table 2 presents some common (crude) types of inference components which, according to our understanding, were included in the various systems (see (Bar-Haim, Szpektor, and Glickman, 2005) and (Vanderwende, Coughlin, and Dolan, 2005) who propose related breakdowns of inference types).

The most basic type of inference measures the degree of word overlap between T and H , possibly including stemming, lemmatization, part of speech tagging, and applying a statistical word weighting such as *inverse document frequency* (idf). Interestingly, a non-participating system that operated solely at this level, using a simple decision tree trained on the development set, obtained an accuracy level of 58%, which might reflect a knowledge-poor baseline (Corley and Mihalcea, 2005). Higher levels of lexical inference considered relationships between words that may reflect entailment, based either on statistical methods or WordNet. Next, some systems measured the degree of match between the syntactic structures of T and H , based on some distance criteria. Finally, few systems incorporated some form of “world knowledge”, and a few more applied a logical prover for making the entailment inference, typically over semantically enriched representations. Different decision mechanisms were applied over the above types of knowledge, including probabilistic models, probabilistic machine translation models, supervised learning methods, logical inference and various specific scoring mechanisms.

Table 3.2 shows the results for the runs as submitted to the challenge. Overall system accuracies were between 50 and 60 percent. Figure 3.1 shows the various accuracy results along with the baseline significance levels of random classification.

Unlike other system submissions, (Vanderwende, Coughlin, and Dolan, 2005) report an interesting manual analysis of the test examples. Each example was analyzed as whether it could be classified correctly (as either true or false) by taking into account only syntactic considerations, optionally augmented by a lexical thesaurus.

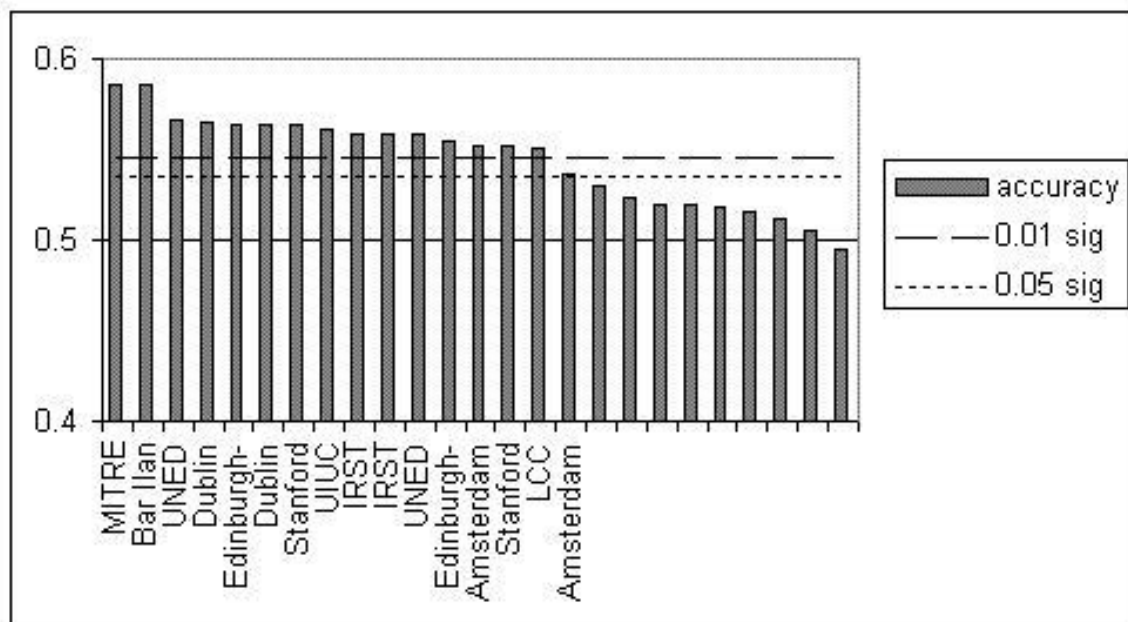


Figure 3.1: Accuracy results for RTE submissions

An “ideal” decision mechanism that is based solely on these levels of inference was assumed. Their analysis shows that 37% of the examples could (in principle) be classified correctly by considering syntax alone, and 49% if a thesaurus is also consulted.

The Comparable Documents (CD) task stands out when observing the performance of the various systems broken down by tasks (see Figure 3.2). Generally the results on this task are significantly higher than results on the other tasks with results as high as 87% accuracy and cws of 0.95. This behavior might indicate that in comparable documents there is a high prior probability that seemingly matching sentences indeed convey the same meanings. We also note that for some systems it is the success on this task which pulled the figures up from the insignificance baselines.

Since the RTE evaluation measures did not favor specifically recognition of positive entailment, a system which does well in recognizing when entailment does not hold did just as well, in terms of accuracy, as a system tailored to recognize true examples.

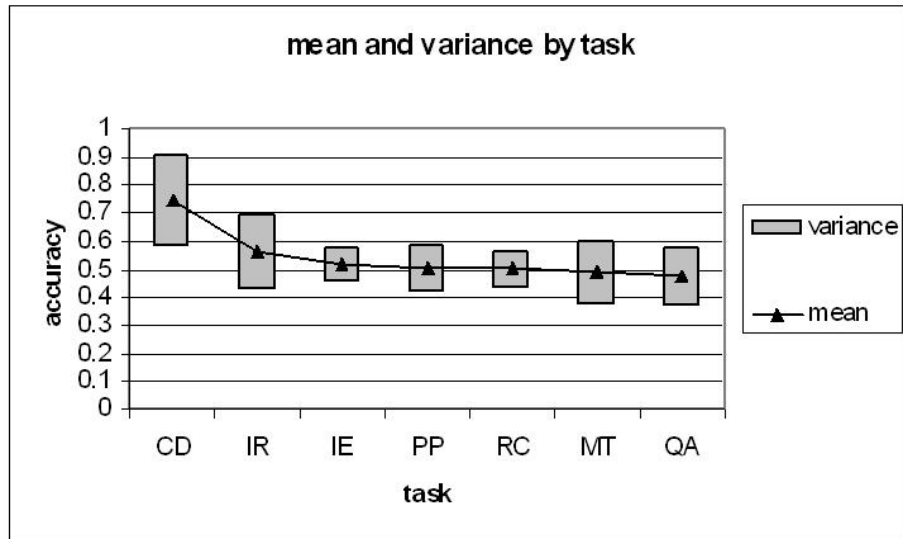


Figure 3.2: Mean and variance of accuracy by task

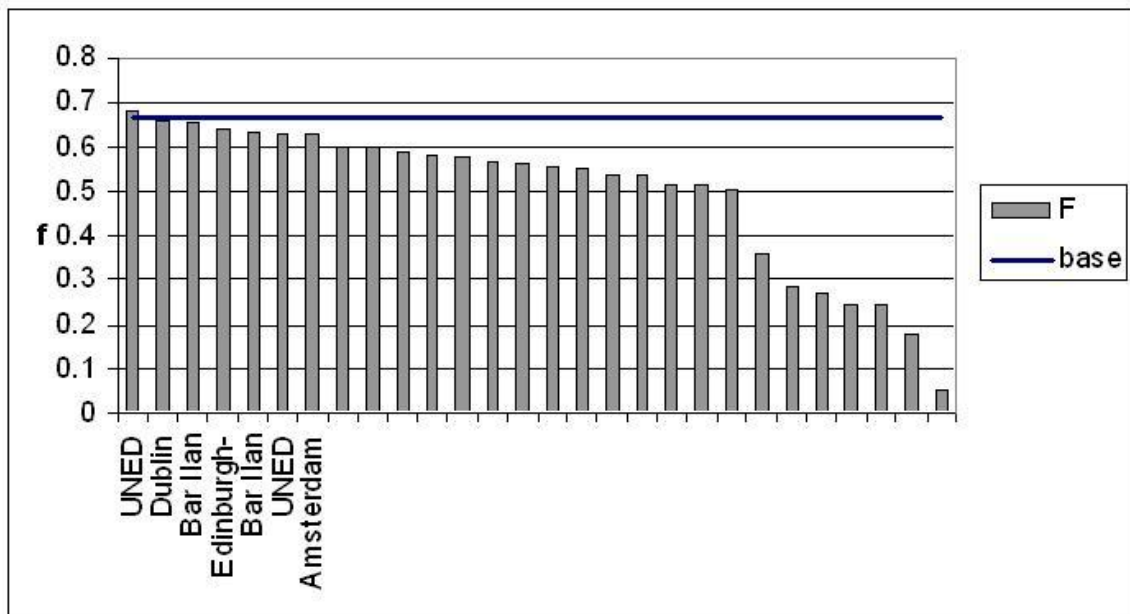


Figure 3.3: F-measure results for the RTE submissions

In fact, some systems recognized only very few positive entailments (a recall between 10-30 percent). In terms of the f-measure (Van Rijsbergen, 1979), none of the systems performed significantly better than the $f=0.67$ baseline of a system which uniformly predicts true (see Figure 3.3). As can be seen, the overall results (including those described in this thesis) are quite low which is not surprising for a composite task of this nature. The results obtained by the participating systems may be viewed as typical for a new and relatively difficult task. Should be regarded as baselines for future more advanced models.

3.5 The Lexical Reference Subtask

3.5.1 Motivation

Decomposing the complex task of entailment into subtasks, and analyzing the contribution of individual NLP components for these subtasks makes a step towards better understanding of the problem, and for pursuing better entailment engines.

An important sub task of textual entailment is recognizing whether a certain textual concept is referenced in a given text. We term this relationship *textual reference* and define it as follows for lexical items:

Definition 2 *A word w is **lexically referred** in a text t if there is an explicit or implied reference in t to a concept denoted by w .*

Textual reference is a natural extension of textual entailment for sub-sentential hypotheses such as words. A concrete version of detailed annotation guidelines for lexical reference is presented in the next section¹.

¹These terms should not be confused with the use of *lexical entailment* in WordNet which describes an entailment relationship between verb lexical types (Miller, 1995) nor with the related notion of *reference* in linguistics generally describing the relation between nouns or pronouns and objects that are named by them (Frege, 1892; Russell, 1919)

It is typically a necessary, but not sufficient, condition for textual entailment that the lexical concepts in a hypothesis h are referred in a given text t . For example, in order to infer from a text the hypothesis “*a dog bit a man*,” it is a necessary that the concepts of *dog*, *bite* and *man* must be referenced in the text, either directly or in an implied manner. However, for proper entailment it is further needed that the right relations would hold between these concepts¹. Furthermore, an entailed hypothesis might contain lexical concepts which are not lexically referred in a text but yet entailment at a whole holds (see following data analysis section 3.5.3). Textual entailment may thus be at best approximated in terms of lexical reference.

3.5.2 Dataset Creation and Annotation Process

We created a lexical reference dataset derived from the RTE development set. We randomly chose 400 out of the 567 text-hypothesis examples of the RTE development set. We then created text-word examples for all content words in the hypotheses which do not appear verbatim in the corresponding text. This resulted in a total of 987 lexical reference examples.

Annotation guidelines

We asked an annotator to annotate the text-word examples according to the following guidelines. Given a text and a target word the annotators were asked to decide whether the target word is referred to in the text or not. They were asked to mark the first applicable criterion from the following list:

morph if a morphological inflection (**infl**), morphological derivation (**deri**) or alternative spelling or some **other** form of the target word is present in the text (see examples 1 (infl), 2 (deri) and 3-5 (other) in Table 3.3).

¹Quoting the known journalism saying – “*Dog bites man*” isn’t news, but “*Man Bites Dog*” is!

trigger if there is a word in the sentence which, in the context of the sentence, shares the same meaning as the target word (synonym), or which implies a reference to the target word's meaning/concept (e.g. hyponym). See examples 6-10 in Table 3.3 where the trigger word is emphasized in the text. Note that in example 10 murder is not a synonym of died nor does it share the same meaning of died; however it is clear from its presence in the sentence that it refers to a death. Also note that in example 19 although home is a possible synonym for house, in the context of the text it does **not** appear in that sense and the example should be annotated as false.

phrase if there is a lexical or lexico-syntactic phrase present in the text that suggests or means the same as the target word, but each word on its own would not constitute a trigger word. See examples 11-13 in Table 3.3.

context in cases where there is a clear reference to the notion of the target word, but there is no specific word or phrase present in the sentence to derive the reference. A reference is rather derived from the general meaning and context of the sentence as a whole. See examples 14-18 in Table 3.3.

false Otherwise, if you find that the target word meaning is not referenced in the sentence in any way, the examples should be annotated as false. In example 20 in Table 3.3 the target word "hiv-positive" should be considered as one word that cannot be broken down from its unit and although both the general term "HIV status" and the more specific term "HIV negative" are referred to, the target word cannot be understood or derived from the text. In example 21 although the year 1945 may refer to a specific war, there is no "war" either specifically or generally understood by the text. And in example 22 although the text refers to a "declaration of independence from the Soviet Union" and uses the phrase "changed hands", one can not specifically infer the notion of 'collapse' from the

phrase nor from the sentence at whole. In these cases, therefore, the example should be annotated as false.

A subset of 400 lexical reference examples were given to a second annotator for cross annotation. In terms of agreement the resulting Kappa value between the five categories was 0.58. When regarding the lexical reference binary task in which the morph, trigger, phrase and context are conflated to true the resulting slightly higher kappa of 0.61 is regarded as substantial agreement (Landis and Koch, 1997).

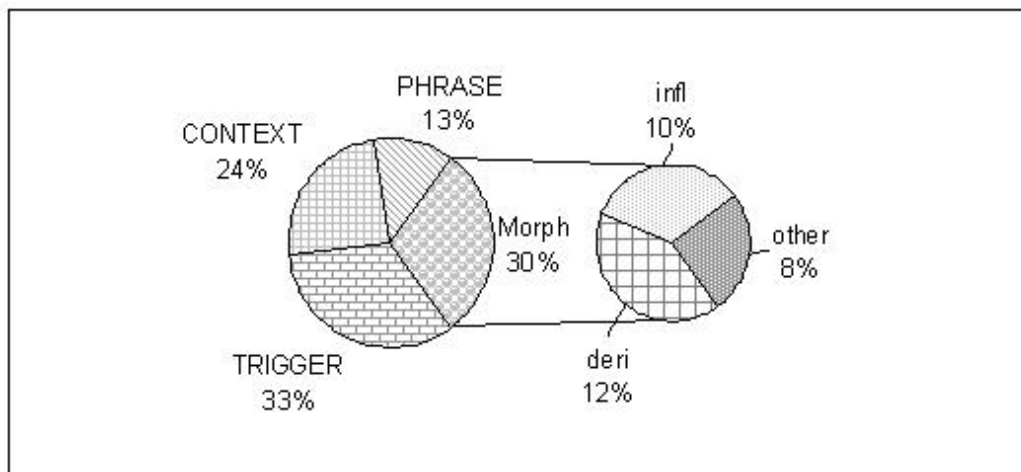


Figure 3.4: The breakdown of the lexical entailment dataset by type out of the 63% of the examples that were judged as true

The performance of a lexical entailment system on this derived dataset can be evaluated via the same evaluation metrics as proposed in Section 3.3. Figure 3.4 shows the breakdown of the lexical entailment dataset by type. Given that the dataset is not balanced in terms of true and false examples, a straw-baseline for accuracy is thus 0.63 representing a system which predicts all examples as true. An average precision greater than 0.66 is better than chance at the 0.05 level and an average precision greater than 0.671 is significant at the 0.01 level.

ID	TEXT	TARGET	VALUE
1	breastfeeding should be promoted for six months	breastfeed	morph
2	The Polish government made Warsaw its capital	Poland	morph
3	George W Bush took office three years ago.	w.	morph
4	China's 90-year-old leader is near death	old	morph
5	Mohammad Said Al-Sahaf, met with the ambassadors	mohammed	morph
6	Mexico's financial services are concentrated in the capital.	centered	trigger
7	Oracle had fought to keep the forms from being released	document	trigger
8	With 549 million in cash, Google can afford to make amends	company	trigger
9	Guerrillas killed a peasant in the city of Flores	civilian	trigger
10	The court found two men guilty of murdering Shapour Bakhtiar	died	trigger
11	The new information prompted them to call off the search	cancelled	phrase
12	They were found with their throats cut in August 1991	died	phrase
13	Milan, home of the famed La Scala opera house, . . .	located	phrase
14	For women who are HIV negative or who do not know their HIV status, breastfeeding should be protected, promoted and supported for six months.	mothers	context
15	Successful plaintiffs recovered punitive damages in Texas discrimination cases 53% of the time.	legal	context
16	Two bombs planted near an Islamic school in Pakistan killed eight people and injured 42	attack	context
17	Recreational marijuana smokers are no more likely to develop oral cancer than nonusers.	risk	context
18	About half were along a 20-mile stretch of Santa Monica Bay from Topanga Canyon Boulevard to the Palos Verdes Peninsula.	coastline	context
19	Pakistani officials announced that two South African men in their custody had confessed to planning attacks at popular tourist spots in their home country.	house	false
20	For women who are HIV negative or who do not know their HIV status, breastfeeding should be promoted for six months.	hiv-positive	false
21	On Feb. 1, 1945, the Polish government made Warsaw its capital, and an office for urban reconstruction was set up.	war	false
22	The Republic of Tajikistan, which declared independence from the Soviet Union in 1991, is in a region that has changed hands many times.	collapsed	false

Table 3.3: Lexical Reference Annotation Examples

3.5.3 Data Analysis

In a similar manner to (Bar-Haim, Szpecktor, and Glickman, 2005; Vanderwende, Coughlin, and Dolan, 2005) we investigated the relationship between lexical reference and textual entailment. We checked the performance of a hypothetical textual entailment system which relies solely on a lexical reference component and asserts that a hypothesis is entailed from a text if and only if all content words in the hypothesis are referred in the text. Based on the lexical entailment dataset annotations and a sample from RTE development set, such an “ideal” system would obtain an accuracy of 69% on the textual entailment task. The corresponding precision is 67% and a recall of 69%. This is significantly higher than the best performing systems that participated in the challenge. Interestingly, a similar entailment system based on a lexical reference component which doesn’t account for the contextual lexical reference would achieve an accuracy of only 63% while maintaining a similar precision but a lower recall of 45%. This suggests that lexical reference in general and contextual entailment in particular play an important (though not sufficient) role in entailment recognition. Table 3.4 lists examples demonstrating when lexical reference does not correlate with entailment. The first examples (20, 85, 180) are false negative examples, in which entailment holds but yet there is a word in the hypothesis (emphasized) which is not referred in the text. Examples (130, 171) are false positives in which all words in the hypothesis are referred in the text but yet entailment does not hold.

3.6 Discussion

As a new task and a first benchmark, textual entailment recognition is still making its first steps towards becoming a mature discipline within the natural language processing community. We received a lot of feedback from participants in the challenge and other members of the research community, which partly contributed to the design

ID	TEXT	HYPOTHESIS	ENTAILMENT	REFERENCE
20	<i>Eating lots of foods that are a good source of fiber may keep your blood glucose from rising too fast after you eat.</i>	<i>Fiber improves blood sugar control.</i>	true	false
85	<i>The country's largest private employer, Wal-Mart Stores Inc., is being sued by a number of its female employees who claim they were kept out of jobs in management because they are women.</i>	<i>Wal-Mart sued for sexual discrimination</i>	true	false
180	<i>The Securities and Exchange Commission's new rule to beef up the independence of mutual fund boards represents an industry defeat.</i>	<i>The SEC's new rule will give boards independence.</i>	true	false
130	<i>Weinstock painstakingly reviewed dozens of studies for evidence of any link between sunscreen use and either an increase or decrease in melanoma.</i>	<i>skin cancer numbers increase.</i>	false	true
171	<i>The terrorist is suspected of being behind several deadly kidnappings and dozens of suicide attacks in Iraq.</i>	<i>Terrorist kidnaps dozens of Iraqis.</i>	false	true

Table 3.4: examples demonstrating when lexical entailment does not correlate with entailment

of the second challenge (RTE-2)¹. Following are some issues that came up at these discussions.

Multi Valued Annotation

In our setting we used a binary {true, false} annotation - a hypothesis is either entailed from the text or not. An annotation of false was used to denote both cases in which the truth value of the hypothesis is either (most likely) false or unknown given the text. Yet, one might want to distinguish between cases (such as example 1667 in Table 3.1) for which the hypothesis is false given the text and cases (such as example 2097) for which it is unknown whether the hypothesis is True or false. For this reason, a 3-valued annotation scheme ({true, false, unknown}; see (Lukasiewicz, 1970)) was proposed as a possible alternative. Furthermore, given the fuzzy nature of the task, it is not clear whether a 3-valued annotation would suffice and so n-valued annotation or even a fuzzy logic scheme (Zadeh, 1965) may be considered as well. Allowing for a richer annotation scheme may enable to include the currently discarded examples on which there was no agreement amongst the annotators (see Section 3.2.3).

Assumed Background Knowledge

Textual inferences are based on information that is explicitly asserted in the text and often on additional assumed background knowledge not explicitly stated in the text. In our guidelines (see Section 3.1.1) we allowed annotators to assume common knowledge of the news domain. However, it is not clear how to separate out linguistic knowledge from world knowledge, and different annotators might not agree on what constitutes common background knowledge. For example, in example 1586 in Table 3.1 one needs to assume world knowledge regarding Arab states and the Arab

¹<http://www.pascal-network.org/Challenges/RTE2/>

language in order to infer the correctness of the hypothesis from the text. Furthermore, the criteria defining what constitutes acceptable background knowledge may be hypothesis dependant. For example, it is inappropriate to assume as background knowledge that The national language of Yemen is Arabic when judging example 1586, since this is exactly the hypothesis in question. On the other hand, such background knowledge might be assumed when examining the entailment “Grew up in Yemen” \rightarrow “Speaks Arabic”. Overall, there seemed to be a consensus that it is necessary to assume the availability of background knowledge for judging entailment, even though it becomes one of the sources for certain disagreements amongst human annotators.

Common Preprocessing

Textual Entailment systems typically rely on the output of several NLP components prior to performing their inference, such as tokenization, lemmatization, part-of-speech tagging, named entity recognition and syntactic parsing. Since different systems differ in their preprocessing modules it becomes more difficult to compare them. In the next challenge we plan to supply some common pre-processing of the data in order to enable better system comparison and to let participants focus on the inference components.

Entailment Subtasks

Textual entailment recognition is a complex task and systems typically perform multiple sub-tasks. It would therefore be interesting to define and compare performance on specific relevant subtasks. For example, in this chapter and in (Bar-Haim, Szpecktor, and Glickman, 2005) we define lexical and lexical-syntactic entailment subtasks and (Marsi and Krahmer, 2005) define an entailment-alignment subtask. Datasets that are annotated for such subtasks may be created in the future.

Inference Scope

Textual entailment systems need to deal with a wide range of inference types. So far we were interested in rather direct inferences that are based mostly on information in the text and background knowledge. Specialized types of inference, such as temporal reasoning, complex logical inference or arithmetic calculations (see example 1911 from Table 3.1) were typically avoided but may be considered more systematically in the future.

“It is a truth very certain that when it is not in our power to determine what is true we ought to follow what is most probable.”

— Descartes

Chapter 4

A Probabilistic setting for Textual Entailment

4.1 Motivation

Textual entailment indeed captures generically a broad range of inferences that are relevant for multiple applications. As mentioned above, a QA system has to identify texts that entail a hypothesized answer. Given the question “*Does John Speak French?*”, a text that includes the sentence “*John is a fluent French speaker*” entails the suggested answer “*John speaks French.*” In many cases, though, entailment inference is uncertain and has a probabilistic nature. For example, a text that includes the sentence “*John was born in France.*” does not strictly entail the above answer. Yet, it is clear that it does increase substantially the likelihood that the hypothesized answer is true.

example	text	hypothesis
1	<i>John is a French Speaker</i>	<i>John speaks French</i>
2	<i>John was born in France</i>	<i>John speaks French</i>
3	<i>Harry's birthplace is Iowa</i>	<i>Harry was born in Iowa</i>
4	<i>Harry is returning to his Iowa hometown</i>	<i>Harry was born in Iowa</i>

Table 4.1: Example of certain and uncertain inferences

The uncertain nature of textual entailment calls for its explicit modeling in probabilistic terms. We therefore propose a general generative probabilistic setting for textual entailment, which allows a clear formulation of probability spaces and concrete probabilistic models for this task. We suggest that the proposed setting may provide a unifying framework for modeling uncertain semantic inferences from texts.

A common definition of entailment in formal semantics (Chierchia and McConnell-Ginet, 2000) specifies that a text t entails another text h (hypothesis, in our terminology) if h is true in *every* circumstance (possible world) in which t is true. For example, in examples 1 and 3 from Table 4.1 we'd assume humans to agree that the hypothesis is necessarily true in any circumstance for which the text is true. In such intuitive cases, textual entailment may be perceived as being certain, or, taking a probabilistic perspective, as having a probability of 1. In quite many other cases, though, entailment inference is uncertain and has a probabilistic nature. In example 2, the text doesn't contain enough information to infer the hypothesis' truth. And in example 4, the meaning of the word hometown is ambiguous and therefore one cannot infer for certain that the hypothesis is true. In both of these cases there are conceivable circumstances for which the text is true and the hypothesis is false. Yet, it is clear that in both examples, the text does increase substantially the likelihood of the correctness of the hypothesis, which naturally extends the classical notion of certain entailment. Given the text, we expect the probability that the hypothesis is indeed true to be relatively high, and significantly higher than its probability of being

-	text	hypothesis
1	<i>Romano Prodi will meet George Bush in his capacity as president of the European commission</i>	<i>Prodi is the president of the European Commission</i>
2	<i>EuroDisney is 30k east of Paris</i>	<i>EuroDisney is in France</i>
3	<i>John Tyler had twenty children and, still, had time to be President of the United States</i>	<i>John Tyler was the president of the united states</i>
4	<i>Besides working in Spain, Victoria Abril also made films in France</i>	<i>Victoria Abril speaks French</i>

Table 4.2: Examples of various uncertain inferences

true without reading the text. Aiming to model application needs, we suggest that the probability of the hypothesis being true given the text reflects an appropriate confidence score for the correctness of a particular textual inference.

When dealing with textual inferences there are many possible underlying causes for uncertainty. Lexical, syntactic, semantic or pragmatic ambiguities in the text or hypothesis are a common cause of uncertainty. In Table 4.2 example 1, entailment depends on a certain probable resolution of the pronoun his in the text as referring to Romano Prodi. In many cases there is missing background knowledge such as in example 2 where entailment is uncertain unless you know that Paris is in France as well as anything 30k east of Paris. In other cases, such as example 3, the entailment is implied from the text but not necessarily logically entailed. And there are also cases such as example 4 where entailment is probable simply due to statistical correlation.

In the next subsections we propose a concrete generative probabilistic setting that formalizes the notion of truth probabilities in such cases.

4.2 A Generative Probabilistic Setting

Let T denote a space of possible texts, and $t \in T$ a specific text. Let H denote the set of all possible hypotheses. A hypothesis $h \in H$ is a propositional statement

which can be assigned a truth value. It is assumed here that h is represented as a textual statement, but in principle it could also be expressed as a text annotated with additional linguistic information or even as a formula in some propositional language.

A semantic state of affairs is captured by a mapping from H to $\{0=\text{false}, 1=\text{true}\}$, denoted by $w : H \rightarrow \{0, 1\}$, called here *possible world* (following common terminology). A possible world w represents a concrete set of truth value assignments for all possible propositions. Accordingly, W denotes the set of all possible worlds.

We assume a probabilistic generative model for texts and possible worlds. In particular, we assume that texts are generated along with a concrete state of affairs, represented by a possible world. Thus, whenever the source generates a text t , it generates also corresponding hidden truth assignments that constitute a possible world w . The probability distribution of the source, over all possible texts and truth assignments $T \times W$, is assumed to reflect inferences that are based on the generated texts. That is, we assume that the distribution of truth assignments is not bound to reflect the state of affairs in a particular “real” world, but only the inferences about propositions’ truth which are related to the text. The probability for generating a true hypothesis h that is not related at all to the corresponding text is determined by some prior probability $P(h)$. For example, $h=$ “*Paris is the capital of France*” might have a prior smaller than 1 and might be false when the generated text is not related at all to Paris or France. In fact, we may as well assume that the notion of textual entailment is relevant only for hypotheses for which $P(h) < 1$, as otherwise (i.e. for tautologies) there is no need to consider texts that would support h ’s truth. On the other hand, we assume that the probability of h being true (generated within w) would be higher than the prior when the corresponding t does contribute information that supports h ’s truth.

It should be clarified that our model assumes that texts are generated along with a *single* possible world. This does not contradict the common viewpoint in semantics

(Chierchia and McConnell-Ginet, 2000) that a text’s meaning is defined by *all* possible worlds with which the text is consistent. In fact, given our generative assumption, one may perceive that the meaning of a text t corresponds to all worlds w for which $P(t, w) > 0$. Yet, assuming that a generation event yields a single world along with the generated text instance enables us to obtain our desired definition of probabilistic entailment. As an intuitive interpretation, this setting assumes that while a text may be consistent with many possible worlds, a particular text instance is still being generated in the context of a concrete fully specified world.

4.2.1 Probabilistic textual entailment definition

We define two types of events over the probability space for $T \times W$:

- I) For a hypothesis h , we denote as Tr_h the random variable whose value is the truth value assigned to h in a given world. Correspondingly, $Tr_h = 1$ is the event of h being assigned a truth value of 1 (true).
- II) For a text t , we use t itself to denote also the event that the generated text is t (as usual, it is clear from the context whether t denotes the text or the corresponding event).

We say that a text t *probabilistically entails* a hypothesis h (denoted as $t \Rightarrow h$) if t increases the likelihood of h being true, that is, if $P(Tr_h = 1|t) > P(Tr_h = 1)$, or equivalently if the pointwise mutual information, $I(Tr_h = 1, t)$, is greater than 0. Once knowing that $t \Rightarrow h$, $P(Tr_h = 1|t)$ serves as a probabilistic confidence value for h being true given t .

Application settings would typically require that $P(Tr_h = 1|t)$ obtains a high value; otherwise, the text would not be considered sufficiently relevant to support h ’s truth (e.g. a supporting text in QA or IE should entail the extracted information with high confidence). Finally, we do not address explicitly the question whether a

text refutes an hypothesis. This seems to be a natural corollary of our setting. Thus the whole frame work provides a probabilistic interpretation of a three way logic in which the truth of an hypothesis can be entailed, refuted or remained unknown given a text. In an analogous manner to the above definition, a text t probabilistically refutes an hypothesis h if it increases the likelihood of the h being false. Since $P(Tr_h = 1) + P(Tr_h = 0) = 1$, this is equivalent to t decreasing the likelihood of the h being true.

4.3 Model Properties

It is interesting to notice the following properties and implications of our probabilistic setting:

Comparison to logical entailment

Textual entailment is defined as a relationship between texts and propositions whose representation is typically based on text as well, unlike logical entailment which is a relationship between propositions only. Accordingly, textual entailment confidence is conditioned on the actual generation of a text, rather than its truth. For illustration, we would expect that the text “*His father was born in Italy*” would entail the hypothesis “*He was born in Italy*” with high probability - since most people who’s father was born in Italy were also born there. However we expect that the text would actually not probabilistically textually entail the hypothesis since most people for whom it is specifically reported that their father was born in Italy were not born in Italy¹.

¹This seems to be the case when analyzing the results of entering the above text in a web search engine.

Comparison to probabilistic reasoning

We assign probabilities to propositions (hypotheses) in a similar manner to Fuzzy Logic (Zadeh, 1965) and certain probabilistic reasoning approaches (e.g. (Bacchus, 1990; Halpern, 1990)). However, we also assume a generative model of text, similar to probabilistic language models and statistical machine translation, which supplies the needed conditional probability distribution. Furthermore, since our conditioning is on texts rather than propositions we do not assume any specific logic representation language for text meaning, and only assume that textual hypotheses can be assigned truth values.

Additionally, unlike fuzzy logic, truth values within a possible world are binary: a certain proposition either holds or not in a specific generated text/world pair. Uncertainty in our model is a consequence of worlds being hidden, and probabilities denote the likelihood that a certain proposition is actually true given a text.

World knowledge

Our framework does not distinguish between textual entailment inferences that are based on knowledge of language semantics (such as *murdering* \Rightarrow *killing*) and inferences based on domain or world knowledge (such as *live in Paris* \Rightarrow *live in France*). Both are needed in applications and it is not clear where and how to put such a borderline.

Comparison to other generative settings

An important feature of the proposed framework is that for a given text many hypotheses are likely to be true. Consequently, for a given text t , $\sum_h P(Tr_h = 1|t)$ does not sum to 1. This differs from typical generative settings for IR and MT (e.g. (Brown et al., 1993; Ponte and Croft, 1998)), where all conditioned events

are disjoint by construction. In the proposed model, it is rather the case that $P(Tr_h = 1|t) + P(Tr_h = 0|t) = 1$, as we are interested in the probability that a single particular hypothesis is true (or false).

Probability estimates

An implemented model that corresponds to our probabilistic setting is expected to produce an estimate for $P(Tr_h = 1|t)$. This estimate is expected to reflect all probabilistic aspects involved in the modeling, including inherent uncertainty of the entailment inference itself (as in example 2 of Table 4.1), possible uncertainty regarding the correct disambiguation of the text (example 4), as well as uncertain probabilistic estimates that stem from the particular model structure and implementation.

*“Improbable as it is, all other explanations
are more improbable still.”*

— Sherlock Holmes

Chapter 5

Lexical Models for Textual Entailment

5.1 Introduction

We suggest that the proposed setting (chapter 4) provides the necessary grounding for probabilistic modeling of textual entailment. Since modeling the full extent of the textual entailment problem is clearly a long term research goal, in this thesis we rather focus on the proposed subtask of *lexical reference* - identifying when the lexical elements of a textual hypothesis are referenced in a given text (see Section 3.5). As described in Section 3.5.3, the textual entailment task can be approximated in terms of lexical reference. Furthermore, we believe that a lexical reference system is an important component in robust textual entailment systems. To model textual entailment at the lexical level we assume that the meaning of each individual content word u in a hypothesis is assigned truth values which correspond to lexical reference. We evaluate our models both on the entailment task itself based on performance on

the RTE dataset (Section 3.3) as well as on the lexical reference subtask based on performance on the derived lexical reference dataset (Section 3.5).

Given this lexically-projected setting, a hypothesis is assumed to be true if and only if all its lexical components are true as well. This captures our target perspective of lexical entailment, while not modeling here other entailment aspects. When estimating the entailment probability we assume first that the truth probability of a term u in a hypothesis h is independent of the truth of the other terms in h , obtaining:

$$\begin{aligned} P(Tr_h = 1|t) &= \prod_{u \in h} P(Tr_u = 1|t) \\ P(Tr_h = 1) &= \prod_{u \in h} P(Tr_u = 1) \end{aligned} \tag{5.1}$$

At this point we describe two different models that estimate the probabilities $P(Tr_u = 1)$ and $P(Tr_u = 1|t)$ for any given word u and text t . The *alignment model* (Section 5.2) assumes an underlying alignment between the text and the hypothesis terms. The *Bayesian model* (Section 5.3) views lexical reference as a text classification task in which entailment is derived from the entire context of the sentence (rather than word-to-word alignment) and Naïve Bayes classification is applied in an unsupervised setting to estimate the hidden reference values. We evaluate and compare the two models on the RTE textual entailment dataset as well as on the lexical reference subtask. We show that both models perform comparably on the textual entailment dataset. However, the strength of the Bayesian model is evident when evaluating on the lexical reference subtask on which it performs significantly better than the alignment model.

5.2 Alignment Model

In this model, in order to estimate $P(Tr_u = 1|t)$ for a given word u and text $t = \{v_1, \dots, v_n\}$, we further assume that the majority of the probability mass comes from

a specific entailing word in t , allowing the following approximation:

$$P(Tr_u = 1|t) = \max_{v \in t} P(Tr_u = 1|T_v) \quad (5.2)$$

where T_v denotes the event that a generated text contains the word v . This corresponds to expecting that each word in h will be entailed from a specific word in t (rather than from the accumulative context of t as a whole). One can view Equation 5.2 as inducing an alignment between terms in the hypothesis and terms in the text, somewhat similar to alignment models in statistical MT (e.g. (Brown et al., 1993)).

Thus we obtain an estimate for the entailment probability based on lexical entailment probabilities from (5.1) and (5.2) as follows:

$$P(Tr_h = 1|t) = \prod_{u \in h} \max_{v \in t} P(Tr_u = 1|T_v) \quad (5.3)$$

We perform unsupervised empirical estimation of the lexical entailment probabilities, $P(Tr_u = 1|T_v)$, based on word co-occurrence frequencies from a given corpus. Following our proposed probabilistic model (cf. Section 4.2), we assume that the corpus is a sample generated by a language source. Each document represents a generated text and a (hidden) possible world. Given that the possible world of the text is not observed we do not know for an observed text the hidden reference values of various words. We therefore further make the simplest assumption that all words appearing verbatim in a text are true in the corresponding world (referenced) and all others are false and hence $P(Tr_u = 1|T_v) \approx P(T_u|T_v)$, the probability that u occurs in a text given that v occurs in that text. The lexical entailment probability estimate

is thus derived from (5.3) as follows:

$$P(Tr_h = 1|t) \approx \prod_{u \in h} \max_{v \in t} P(T_u|T_v) \quad (5.4)$$

The co-occurrence probabilities are easily estimated based on maximum likelihood counts:

$$P(T_u|T_v) = \frac{n_{u,v}}{n_v} \quad (5.5)$$

where n_v is the number of documents containing word v and $n_{u,v}$ is the number of documents containing both u and v . In the experiments we investigated obtaining the corresponding counts from two sources. First by performing queries to a web search engine, since the majority of RTE examples were based on web snippets, and the second from a large corpus in the news domain.

5.2.1 Web-based Estimation of Lexical Entailment Probabilities

The text and hypothesis of all pairs in the RTE development and test sets were tokenized by the following simple heuristic - split at white space and remove any preceding or trailing of the following punctuation characters: ([{}])”“.,;:-!?. A standard stop word list was applied to remove frequent tokens. Counts were obtained using the AltaVista search engine¹, which supplies an estimate for the number of results (web-pages) for a given one or two token query.

We empirically tuned a threshold, λ , on the estimated entailment probability to decide if entailment holds or not. For a $t - h$ pair, we tagged an example as true (i.e. entailment holds) if $p = P(Tr_h = 1|t) > \lambda$, and as false otherwise. We assigned a confidence of p to the positive examples ($p > \lambda$) and a confidence of $1 - p$ to the

¹<http://www.av.com/>

task	accuracy
Comparable Documents (CD)	0.8333
Machine Translation (MT)	0.5667
Information Extraction (IE)	0.5583
Reading Comprehension (RC)	0.5286
Paraphrase (PP)	0.5200
Information Retrieval (IR)	0.5000
Question Answering (QA)	0.4923
Overall	0.586

Table 5.1: Accuracy results by task

negative ones.

The threshold was tuned on the 567 annotated text-hypothesis example pairs in the development. The optimal threshold of $\lambda = 0.005$ resulted in an accuracy of 56% on the development set. This threshold was used to tag and assign confidence scores to the 800 pairs of the test set.

Analysis and Results

The resulting overall accuracy on the test set was of 59% which is statistically significantly better than chance at the 0.01 level. This system competed at the RTE challenge and tied for first place in terms of accuracy (see Table 3.2, page 36). Table 5.1 lists the accuracy when computed separately for each task in the RTE dataset. As can be seen by the table the system does well on the CD and MT tasks, and quite poorly (not significantly better than chance) on the RC, PP, IR and QA tasks. It seems as if the success of the system is attributed almost solely to its success on the CD and MT tasks. Indeed it seems as if there is something common to these two tasks, which differentiates them from the others - in both tasks high correspondence at the lexical level tends to correlate with entailment.

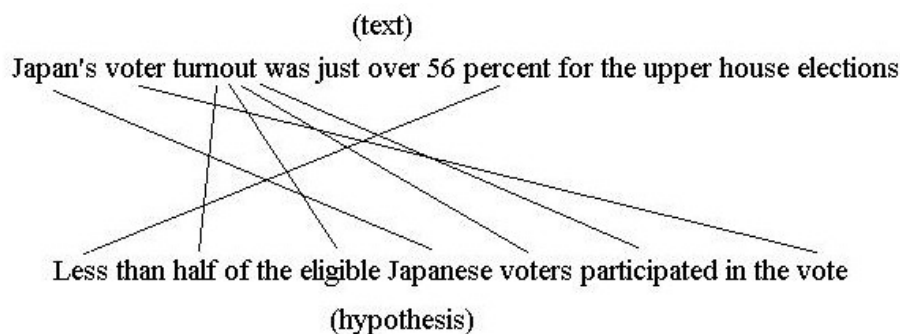


Figure 5.1: System’s underlying alignment for example 1026 (RC). gold standard - false, system - false

Success and failure cases

The system misclassified 331 out of the 800 test examples. The vast majority of these mistakes (75%) were false positives - pairs the system classified as true but were annotated as false. It is also interesting to note that the false negative errors were more common among the MT and QA tasks while the false positive errors were more typical to the other tasks. An additional observation from the recall-precision curve (Figure 5.3) is that high system confidence actually corresponds to false entailment. This is attributed to an artifact of this particular dataset by which examples with high word overlap between the text and hypothesis tend to be biased to negative examples (see Section 3.2.1).

In an attempt to ‘look under the hood’ we examined the underlying alignment obtained by our system on a sample of examples. Figure 5.1 illustrates a typical alignment. Though some of the entailing words correspond to what we believe to be the correct alignment (e.g. voter → vote, Japan’s → Japanese), the system also finds many dubious lexical pairs (e.g. turnout → half, percent → less). Furthermore, some of the induced alignments do not correspond to the “expected” ones. For example, in Figure 5.2 - based on the web co-occurrence statistics, *detonated* is a better trigger

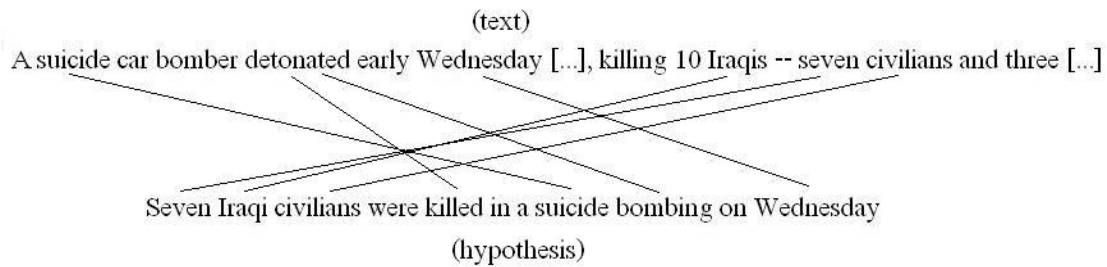


Figure 5.2: System’s underlying alignment for example 1095 (202). gold standard - true, system - true

word for both *killed* and *bombing* even though one would expect to align them with the words *killing* and *bomber* respectively. Obviously, co-occurrence within documents is only one factor in estimating the entailment between words. This information should be combined with other statistical criteria that capture complementary notions of entailment, such as lexical distributional evidence as addressed in Chapter 6, in (Geffet and Dagan, 2004; Geffet and Dagan, 2005), or with lexical resources such as WordNet (Miller, 1995).

Comparison to baseline

As a baseline model for comparison we use a heuristic score proposed within the context of text summarization and Question Answering (Monz and de Rijke, 2001; Saggion et al., 2004) (see Section 2.2). In this score semantic overlap between two texts is modeled via a word overlap measure, considering only words that appear in both texts weighted by *inverse document frequency* (*idf*). More concretely, this directional entailment score between two texts, denoted here by $entscore(t, h)$, is defined as follows:

$$entscore(t, h) = \frac{\sum_{w \in t \cap h} idf(w)}{\sum_{w \in h} idf(w)} \quad (5.6)$$

model	average precision
<i>align</i>	0.526
<i>entscore₂</i>	0.525
<i>entscore</i>	0.523

Table 5.2: Average precision results for alignment and baseline models

where $idf(w) = \log(N/n_w)$, N is the total number of documents in the corpus and n_w the number of documents containing word w . We have tested the performance of this measure in predicting entailment on the RTE dataset. Tuning the classification threshold on the development set (as done for our system), *entscore* obtained a somewhat lower accuracy of 56%.

To further investigate the contribution of the co-occurrence probabilities we extended the *entscore* measure by incorporating lexical co-occurrence probabilities in a somewhat analogous way to their utilization in our model. In this extended measure, termed *entscore₂*, we compute a weighted average of the lexical probabilities, rather than their product in our model (Equation 5.3), where the weights are the *idf* values, following the rational of the *entscore* measure. More concretely, *entscore₂* is defined as follows:

$$entscore_2(t, h) = \frac{\sum_{u \in h} idf(u) * \max_{v \in t} P(Tr_u = 1|v)}{\sum_{u \in h} idf(u)} \quad (5.7)$$

$P(Tr_u = 1|v)$ is approximated by $P(T_u|T_v)$ and estimated via co-occurrence counts, as in our model (equations 5.4 and 5.5). Note that when using this approximation, $P(Tr_u = 1|v) = 1$ when $u = v$ and thus the max value in (5.7) is obtained as 1 for hypothesis words that appear also in the text, naturally extending the rational of *entscore*.

Figure 5.3 compares the recall-precision curves for our system and the two baseline entailment scores. The different recall points are obtained by varying a threshold over the entailment score (or probability), considering all examples with a score higher

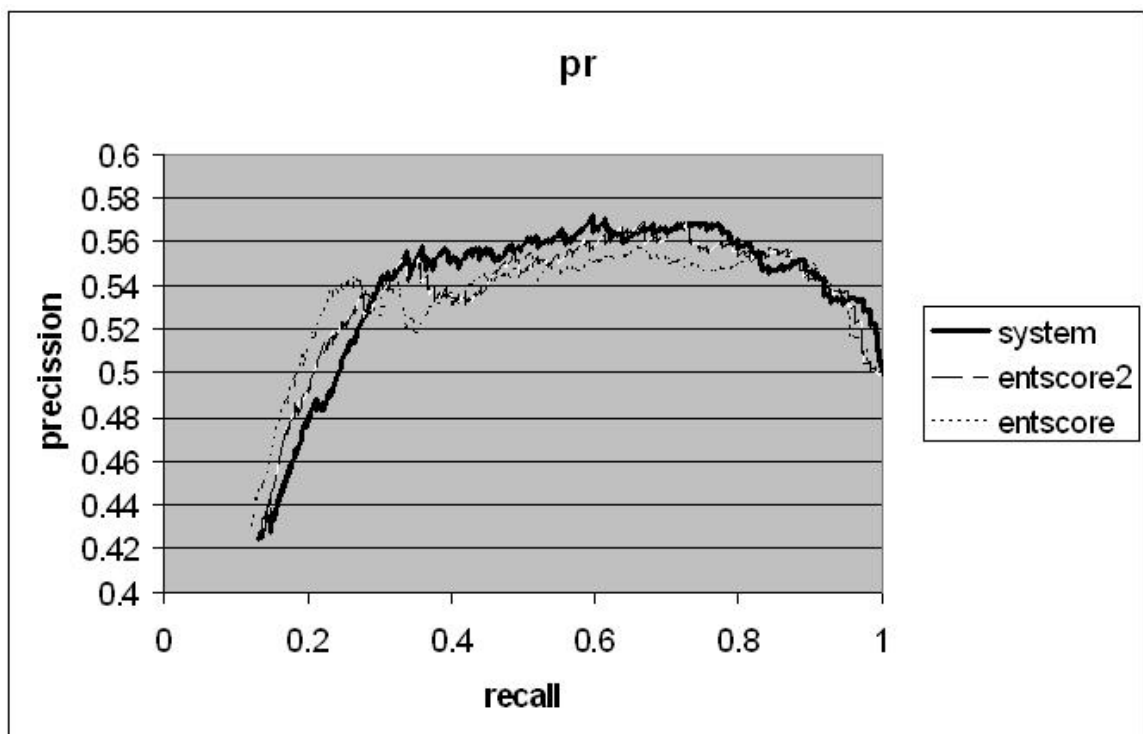


Figure 5.3: Comparison to baselines (*system* refers to our probabilistic model)

level	accuracy	average precision
sentence	0.571	0.515
paragraph	0.572	0.515
document	0.576	0.514
web	0.585	0.526

Table 5.3: Accuracy and average precision for the various co-occurrence levels.

than the threshold as positive classifications. The figures show that on this dataset our system has higher precision over most recall ranges¹. In addition, *entscore₂*, which incorporates lexical co-occurrence probabilities, performs somewhat better than the baseline *entscore* which considers only literal lexical overlap. This behavior is also illustrated by the average precision results shown in Table 5.2. These results demonstrate a marginal contribution of (i) utilizing lexical co-occurrence probabilities and (ii) embedding them within a principled probabilistic model.

5.2.2 Corpus-based Estimation of Lexical Entailment Probabilities

In an additional experiment we used co-occurrence statistics from a news corpus rather than from the web. Experiments were done on the Reuters Corpus Volume 1 (Rose, Stevenson, and Whitehead, 2002) - a collection of about 810,000 English News stories most of which are economy related.

When working with a corpus rather than web-counts one is not restricted to document co-occurrence statistics. We experimented with estimating the lexical entailment probabilities of (5.5) based on document, paragraph and sentence co-occurrence counts within the corpus. Figure 5.4 show the recall-precision curves on the RTE test

¹Note the anomaly that high lexical overlap, which yields high system confidence, actually correlates with false entailment (as noted in Section 3.2.1). This anomaly explains the poor precision of all systems at the lower recall ranges, while the generally more accurate models are effected more strongly by this anomaly.

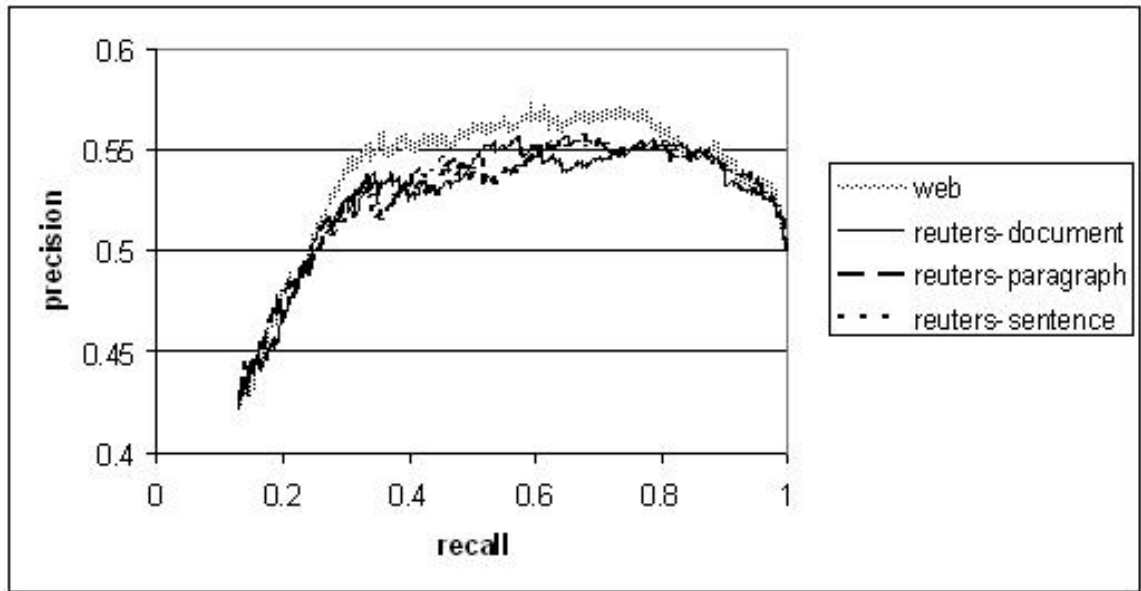


Figure 5.4: Recall-Precision on the RTE test using co-occurrence counts at various levels from the Reuters corpus

set using co-occurrence counts at various levels from the Reuters corpus. Similarly, Table 5.3 lists the accuracy and average precision scores. The relatively low average precision scores reaffirm that high system confidence actually correlates with false entailment (see Section 3.2.1). As can be seen from the graph and table, using counts from the web outperforms using counts from the corpus. In addition, there is no significant difference in performance between using co-occurrence counts from Reuters document, paragraph and sentence levels.

The RTE data was created during 2005 and was based on many news stories and events from that year. The Reuters corpus, however, consists of news articles from 1996. As it turns out, roughly 4% of the words in the RTE data set do not appear at all in the Reuters corpus. Thus it is not surprising that a model based on a large but yet almost a decade old corpus is outperformed by one based on the huge and up-to-date web. Looking at the list of RTE words which do not appear in the Reuters corpus,

one can find many Iraq related words (e.g. Fallujah, Falluja, Baquba, Baqubah, al-Qaeda, Zarqawi), many internet and hi-tec related words (e.g. eBay, Google, iTunes, Napster, MySQL, internet-borne) and ones related to events from the past years (e.g. Britney, Clinton-Monica, Putin, SpaceShipOne).

5.2.3 Performance on the Lexical Reference Subtask

Given that our models were targeted to model lexical entailment rather than the complete entailment task, we also evaluated the performance of the alignment model on the lexical reference dataset (see Section 3.5). Figure 5.5 shows average precision

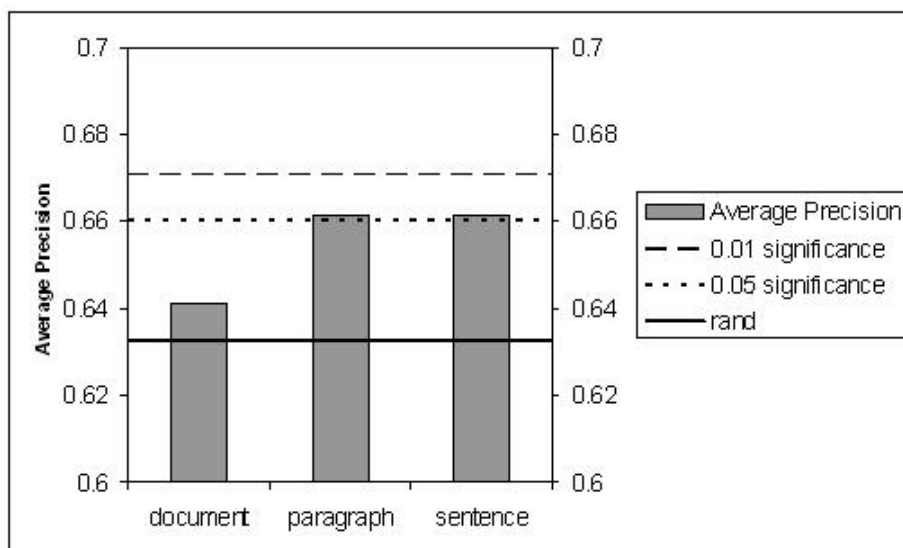


Figure 5.5: Lexical Reference average precision results for alignment model based on corpus co-occurrence counts at different co-occurrence levels

results for alignment model based on corpus co-occurrence counts at different co-occurrence levels on the lexical reference task. Consequently, the alignment model's scores do correspond with lexical reference though only sentence or paragraph co-occurrences are significant (at the 0.05 level). We further analyze the results on the lexical reference dataset in Section 5.3.6.

5.3 Bayesian Model

5.3.1 Introduction

The proposed alignment model, though performing competitively on the RTE dataset, is quite simple and makes several obviously wrong assumptions. In this section we present a model which avoids these assumptions. The model views lexical entailment as a text classification task. Entailment is derived from the entire context of the sentence (rather than from word-to-word alignment) and Naïve Bayes classification is applied in an unsupervised setting to estimate the hidden lexical truth assignments.

5.3.2 Textual Entailment as Text Classification

At this point, it is perhaps best to think of the entailment problem as a text classification task. Our main sleight-of-hand here is estimating these probabilities, $P(Tr_u = 1|t)$ for a text t and a lexical item u as text classification probabilities in which the classes are the different words u in the vocabulary. Following this perspective we apply a technique commonly used for text classification. Our proposed model resembles work done on text classification from labeled and unlabeled examples (Nigam et al., 2000), using a Naïve Bayes classifier. However, our setting and task is quite different - we classify texts to a binary abstract notion of lexical truth rather than to well-defined supervised classes. We utilize unsupervised initial approximate labeling of classes, while Nigam et al. bootstrap from labeled data. First, we construct the initial labeling based solely on the explicit presence or absence of each u in t . Then we apply Naïve Bayes in an unsupervised fashion which is derived from the defined probabilistic setting.

5.3.3 Initial Labeling

As an initial approximation, we assume that for any document in the corpus the truth value corresponding to a term u is determined by the explicit presence or absence of u in that document. Thus, referring to the given corpus texts at the document level, we have $P(Tr_u = 1|t) = 1$ if $u \in t$ and 0 otherwise, which defines the initial class labels for every term u and text t (training labels). It also follows from (5.3) that a text entails a hypothesis if and only if it contains all content words of the hypothesis. In some respect the initial labeling is equivalent to systems that perform a Boolean search (with no expansion) on the keywords of a textual hypothesis in order to find candidate (entailing) texts. Of course, due to the semantic variability of language, similar meanings could be expressed in different wordings, which is addressed in the subsequent model. The initial labeling, however, may provide useful estimates for this model.

5.3.4 Naïve Bayes Refinement

Based on the initial labeling we consider during training all texts that include u as positive examples for this class and take all the other texts as negative examples. For a word u , $P(Tr_u = 1|t)$ can be rewritten, by following the standard Naïve Bayes assumption, as in (5.8):

$$\begin{aligned}
 P(Tr_u = 1|t) &= \frac{P(t|Tr_u=1)P(Tr_u=1)}{P(t)} = && \text{Bayes} \\
 &= \frac{P(t|Tr_u=1)P(Tr_u=1)}{P(t|Tr_u=0)P(Tr_u=0)+P(t|Tr_u=1)P(Tr_u=1)} = && \\
 &= \frac{P(Tr_u=1) \prod_{v \in t} P(v|Tr_u=1)^{n(v,t)}}{P(Tr_u=0) \prod_{v \in t} P(v|Tr_u=0)^{n(v,t)} + P(Tr_u=1) \prod_{v \in t} P(v|Tr_u=1)^{n(v,t)}} = && \text{independence} \\
 &= \frac{P(Tr_u=1) \prod_{v \in V} P(v|Tr_u=1)^{n(v,t)+\epsilon}}{P(Tr_u=0) \prod_{v \in V} P(v|Tr_u=0)^{n(v,t)+\epsilon} + P(Tr_u=1) \prod_{v \in V} P(v|Tr_u=1)^{n(v,t)+\epsilon}} = && \text{smoothing}
 \end{aligned} \tag{5.8}$$

where $n(w, t)$ is the number of times word w appears in t , V is the vocabulary

and ϵ is some small smoothing factor. The smoothing is done by adding epsilon-count words to each sentence in the data to avoid zero probabilities. Similarly, we have $P(Tr_u = 0|t) = 1 - P(Tr_u = 1|t)$. This estimation procedure is known as the multinomial event model with Laplace smoothing and is commonly used for text categorization (McCallum and Nigam, 1998).

In this way we are able to estimate $P(Tr_u = 1|t)$ based solely on the prior probabilities $P(Tr_u = 1), P(Tr_u = 0)$ and the lexical co-occurrence probabilities $P(v|Tr_u = 1), P(v|Tr_u = 0)$ for u, v in the vocabulary V . These probabilities are easily estimated from the corpus given the initial model's estimate of truth assignments as in (5.9).

$$\begin{aligned} P(Tr_u = 1) &= \frac{|d : u \in d|}{|d|} \\ P(v|Tr_u = 1) &= \frac{\sum_{d:u \in d} (n(v, d) + \epsilon)}{\sum_{w \in V} \sum_{d:u \in d} (n(w, d) + \epsilon)} \end{aligned} \tag{5.9}$$

From equations (5.8) and (5.9) we have a refined probability estimate for $P(Tr_u = 1|t)$ for any arbitrary text t and word u .

5.3.5 Experimental Setting

Given that in our setting we are interested in entailment from sentences rather than whole documents and given that the sentence level co-occurrence based alignment model of section 5.2.2 performed better than the document level model, we used the corpus sentences as our atomic unit of text in the generative model. We trained our model over the the sentences of the Reuters corpus. The XML tagged corpus comes along with paragraph markers. Additional breaks to the sentence level was done using the MxTerminator sentence boundary detection software (Reynar and Ratnaparkhi,

1997)¹ and sentences were tokenized using the Stanford NLP Group JAVA software toolkit². Tokens were not stemmed nor lemmatized.

In order to reduce the data size we performed feature selection on the data for each model learned. We used the *information gain* measure which is commonly applied for text classification tasks (Yang and Pedersen, 1997; Forman, 2003). For a given target word we chose the top 5000 most informative keywords and filtered any token which appeared in less than 100 sentences. As a smoothing factor we chose $\epsilon = 0.001$. These parameters were based on ad hoc tuning on a set of held out keywords prior to experimentation.

Table 5.4 demonstrates the lexical entailment probability estimation process for the target word ‘job’. It contains a snapshot of the corpus sentences – some containing the target word (initially labeled as 1) and some not (initially labeled with lexical entailment probability of 0). P_1 denotes the actual resulting lexical entailment probability estimates. Note that for sentences containing the target word, the initial labeling is not changed and remains 1 even though the model might assign them a probability estimate less than 1. Informative words in the text are emphasized. Table 5.5 lists the top trigger words for *job* and $\neg job$.

5.3.6 Empirical Results

Figure 5.6 compares the recall-precision curves of the alignment and Bayesian models on the RTE dataset based on Reuters’ sentence level co-occurrence statistics. As can be seen, the models are comparable. Also the average precision (0.511) and accuracy (0.565) are comparable to those of the alignment model (cf. Table 5.3). However, the difference between the models is evident when comparing them on the lexical entailment dataset. As can be seen in the recall-precision curves of Figure 5.7, the Bayesian

¹<http://www.cis.upenn.edu/~adwait/>

²<http://nlp.stanford.edu/software/index.shtml>

id	text	P ₀	P ₁
5210.8.0	“So you can obviously expect a lot of focus on job creation and particularly job creation in inner-city areas where there are concentrations of welfare recipients.”	1	1(0.99)
74768.2.0	The spokesman said all the U.S. company’s larger plants in Britain had targets for job cuts but any job losses would be voluntary and the company had the full co-operation of unions .	1	1(0.99)
77291.2.0	- To many Finnish voters in the coming local elections, an ideal local government representative is well educated, supports job creation and defends womens’ rights.	1	1(0.78)
42964.2.2	And I don’t think Canadians want us to stop the job when it is only half done.	1	1(0.66)
	...	1	
89214.10.0	According to work council representatives, some 7,000 workers walked out at the Bochum plant of General Motors’ Adam Opel unit, while some 12,000 early-shift workers at Ford’s German plant in Cologne also held a meeting in protest at the cuts .	0	0.92
86239.8.0	The CNPF decision comes the same day as a strike by teachers against cuts of 2,300 in civil service education staff levels next year under the centre-right government’s 1997 spending-cut budget and regional protests by rail workers over jobs .	0	0.83
103280.3.0	The newspaper said Sony would also name John Calley of United Artists Pictures Inc and Sony Corp of America Executive Vice President Jeffrey Sagansky to top management positions within the company , as had been previously announced.	0	0.16
82556.8.0	However, a strong bond market had offset losses to some extent, with the market ending well off its intraday lows.	0	0.01
	...	0	
	...	0	

Table 5.4: The lexical entailment probability estimation process - P₀ represents the initial labeling and P₁ the Bayesian estimation for P($Tr_{job} = 1|t$)

u	$P(Tr_{job} = 1 u)$	v	$P(Tr_{job} = 0 v)$
job	0.517	newsroom	0.99970
cuts	0.122	shares	0.99966
workers	0.104	tonnes	0.99964
jobs	0.097	closed	0.99944
UAW ^a	0.095	cents	0.99944
Unions	0.093	index	0.99943
GM	0.082	traders	0.99935
Ford	0.081	pct	0.99935
Chrysler	0.080	yen	0.99934
labor	0.077	oil	0.99931
strike	0.073	wheat	0.99926
Auto	0.062	trading	0.99926
employment	0.062	bln	0.99925
wage	0.061	U.K.	0.99924
outsourcing	0.055	profit	0.99923
welfare	0.054	net	0.99922
creation	0.054	dollar	0.99921
wages	0.053	futures	0.99919
work	0.050	share	0.99919
spending	0.049	Sept	0.99916
union	0.049	prices	0.99912
employers	0.049	note	0.99910

Table 5.5: Top scoring trigger words for *job* and \neg *job*

^aabbreviation of United Automobile Workers

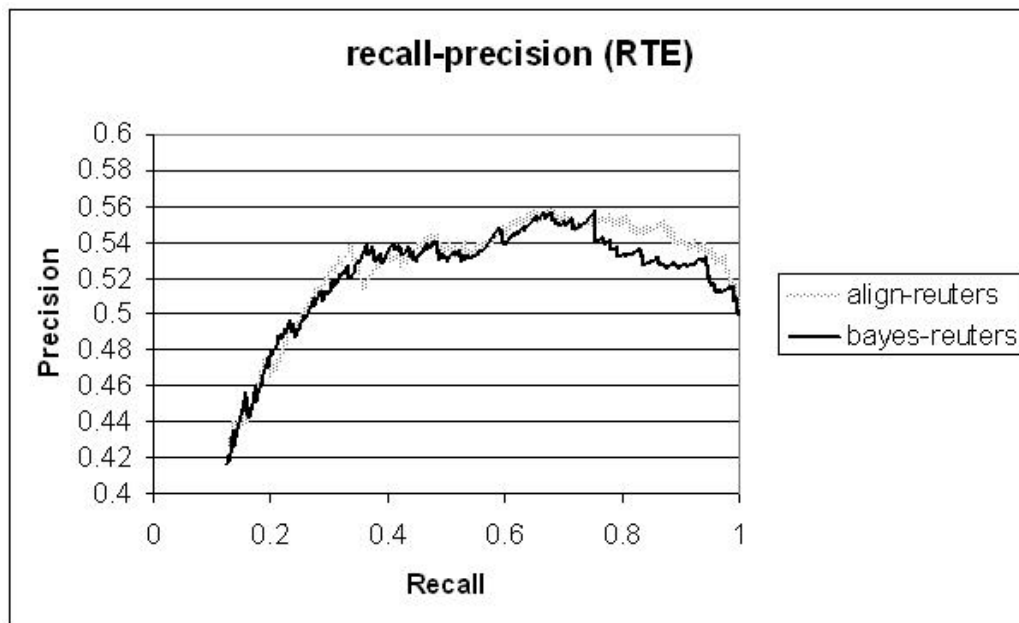


Figure 5.6: Recall Precision comparison on the RTE dataset for Reuters based alignment and bayesian models.

model has substantially higher precision when most confident (the low recall levels). Figure 5.7 also demonstrates that the Bayesian model performs significantly better than the alignment model in terms of average precision. In terms of accuracy none of the models performs better than the 0.63 baseline which results from a threshold of 0 (i.e. assuming all examples are correct).

5.3.7 Running Additional EM Iterations

A natural expected experiment is to apply further Expectation Maximization (EM) iterations to achieve better entailment probability estimates. When applying EM, Equation (5.8) constitutes the E step. One can describe the generative story as follows. For a given target word, u , we assume that our language source generates a sentence of length l , by first choosing the truth value, $c \in \{0, 1\}$, of u based on a probability distribution $P(Tr_u = c)$. It then independently generates the l words of

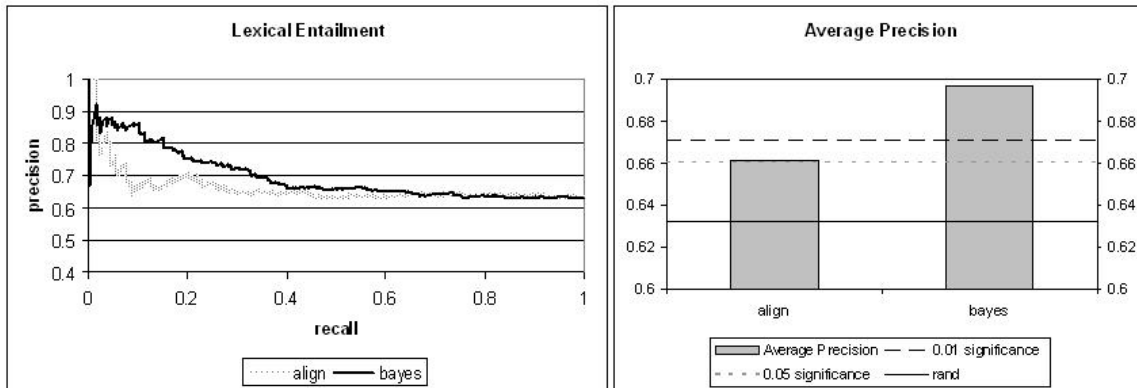


Figure 5.7: comparison of Recall Precision and average precision for Reuters based alignment and bayesian models on the lexical entailment task.

the sentence w_1, \dots, w_l based on a conditional probability distribution $P(w|Tr_u = c)$. We assume that the words of a sentence are generated independently of context and position and are dependant only on the truth of u (The Naïve Bayes assumption). The observed data is a corpus (a set of sentences) $D = (d_1 \dots d_m)$. The unobserved data is the underlying truth value of u , for each sentence $d \in D$ for which $u \notin d$. The likelihood of the data for a given model is:

$$P(D) = \prod_{d \in D} P(d) = \prod_{d \in D} \left[\sum_c P(c) \prod_w P(w|c)^{n(w,d)+\epsilon} \right] \quad (5.10)$$

The m-step is equivalent to Equation (5.9) and is rewritten as follows (5.11):

$$\begin{aligned} P(Tr_u = 1) &= \frac{\sum_d \hat{P}(Tr_u = 1|d)}{|d|} \\ P(v|Tr_u = 1) &= \frac{\sum_d \hat{P}(Tr_u = 1|d) (n(v, d) + \epsilon)}{\sum_{w \in V} \sum_d \hat{P}(Tr_u = 1|d) (n(w, d) + \epsilon)} \end{aligned} \quad (5.11)$$

Where \hat{P} are the probability estimates obtained and updated at each e-step. In a similar manner to (Nigam et al., 2000), for a document d which contains the target word u the truth assignment $P(Tr_u = 1|d)$ is assumed to be known (as 1) and is not

updated during the e-steps. The goal of applying the EM is to estimate the unknown hidden truth assignments when $u \notin d$.

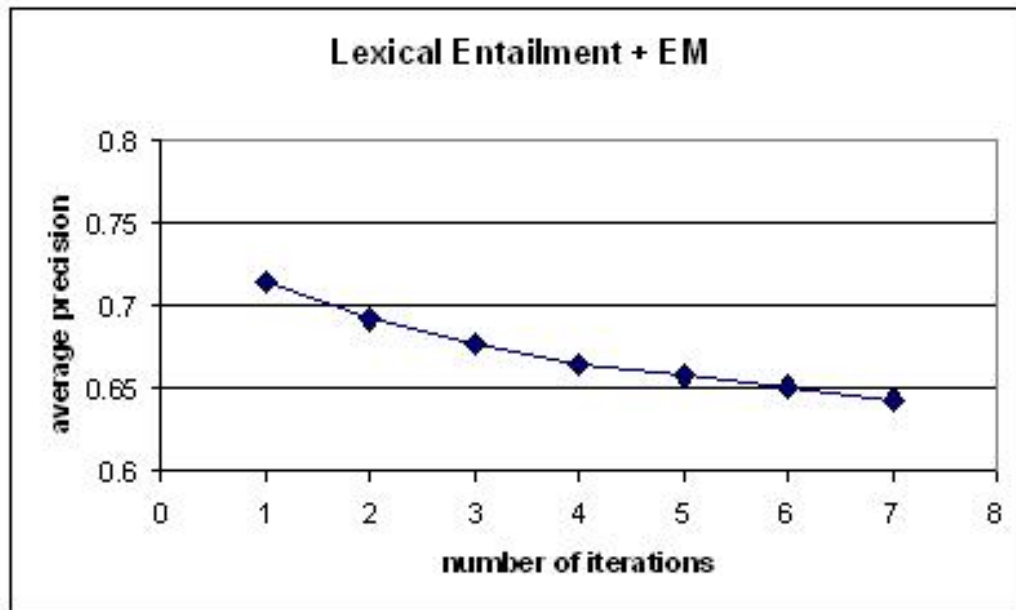


Figure 5.8: Average precision results on lexical entailment dataset for different vs. number of EM iterations

As can be seen in Figure 5.8, additional EM iterations of the Bayesian model do not improve the average precision results on the lexical reference task. It is not atypical that additional EM iterations do not improve results (e.g. (Berger and Lafferty, 1999)) Nevertheless, we are planning to further investigate the reason why EM degrades performance. Devising possible adaptations of the model addressing this issue becomes a challenging target for future research.

5.4 Discussion

Table 5.6 lists a few examples from the lexical reference dataset along with their gold-standard annotation and the Bayesian model score. Manual inspection of the data

shows that the Bayesian model commonly assigns a low score to correct examples which have a morphological variant or entailing trigger word in the sentence but yet the context of the sentence as a whole is not typical for the target hypothesized entailed word. For example, in example 9 the entailing phrase 'set in place' and in example 10 the morphological variant 'founder' do appear in the text however the contexts of the sentences are not typical news domain contexts of issued or founded. An interesting future work would be to change the generative story and model to account for such cases. Consequently, another important research direction is to independently focus on the acquisition of lexical relations which constitute or suggest lexical reference. An initial attempt in this direction is presented in Chapter 6.

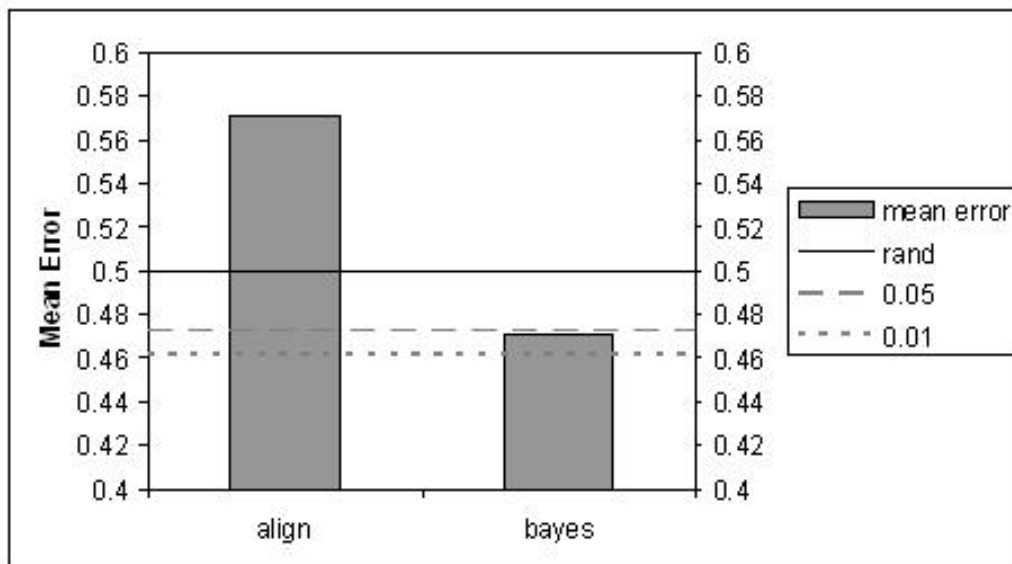


Figure 5.9: Mean error rates for the alignment and Bayesian models on the lexical reference dataset. Note that a lower mean error is better.

In another test we wanted to check the quality of the resulting probability estimates rather than just their ranking ability. Figure 5.9 shows the mean error of the two models on the lexical reference dataset. The mean error is calculated as $meanerror = \frac{\sum_{i=1}^n |P_i - T_i|}{n}$ where i runs over the examples in the dataset, T_i is the

id	text	token	annotation	score
1	<i>QNX Software Systems Ltd., a leading provider of real-time software and services to the embedded computing market, is pleased to announce the appointment of Mr. Sachin Lawande to the position of vice president, engineering services.</i>	named	SYNTAX	0.98
2	<i>NIH's FY05 budget request of \$28.8 billion includes \$2 billion for the National Institute of General Medical Sciences, a 3.4-percent increase, and \$1.1 billion for the National Center for Research Resources, and a 7.2-percent decrease from FY04 levels.</i>	reduced	TRIGGER	0.91
3	<i>Pakistani officials announced that two South African men in their custody had confessed to planning attacks at popular tourist spots in their home country.</i>	security	CONTEXT	0.80
4	<i>With \$549 million in cash as of June 30, Google can easily afford to make amends.</i>	company	TRIGGER	0.18
5	<i>Hepburn, a four-time Academy Award winner, died last June in Connecticut at age 96.</i>	won	MORPH	0.14
6	<i>A senior al-Qaeda operative who was said to be planning an attack on Heathrow Airport has been arrested.</i>	home	FALSE	0.06
7	<i>With \$549 million in cash as of June 30, Google can easily afford to make amends.</i>	shares	FALSE	0.03
8	<i>There are many Baroque churches of the Counter-Reformation period, including the Jesuit Church next to the cathedral and the Church of the Holy Cross, which contains Chopin's heart.</i>	capital	FALSE	0.03
9	<i>In the year 538, Cyrus set in place a policy which demanded the return of the various gods to their proper places.</i>	issued	SYNTAX	7e-4
10	<i>The black Muslim activist said that he had relieved Muhammad of his duties "until he demonstrates that he is willing to conform to the manner of representing Allah and the honorable Elijah Muhammad (founder of the Nation of Islam)".</i>	founded	MORPH	3e-6
11	<i>Budapest is Europe's largest spa town; there are more than 100 hot springs that spout from Buda's limestone bedrock.</i>	properties	FALSE	1e-7

Table 5.6: A sample from the lexical reference dataset along with the Bayesian model's score

gold annotation (true=1, false=0) of example i and P_i is the corresponding model’s probability estimate. The resulting mean error of alignment model is significantly worse than random. However, the Bayesian model has more accurate probability estimates and is significant at the 0.05 level. This further suggests that the Bayesian model is superior to the alignment models in terms of estimating the lexical reference probabilities $P(Tr_u = 1|t)$ for a text t and word u .

The “stronger” Bayesian model which performs better than the alignment model on the lexical reference task doesn’t perform better on the RTE dataset. This could be attributed to the mentioned characteristic of the RTE dataset that high lexical overlap actually correlates with false entailment as well. In addition modeling the overall task of textual entailment involves many aspects beyond the lexical level and lexical reference in particular. Nevertheless, we believe that future and improved textual entailment systems will rely and contain a lexical reference component. For this reason it is important to compare and perfect dedicated lexical reference models by evaluating them directly on the lexical reference subtask they are targeted to model. As a consequence, better lexical reference models will thus set forth, in the long run, improved textual entailment systems.

5.5 Related Work

Different techniques and heuristics were applied on the RTE-1 dataset to specifically model textual entailment. Interestingly, a number of works (e.g. (Bos and Markert, 2005; Corley and Mihalcea, 2005; Jijkoun and de Rijke, 2005)) applied or utilized a lexical based word overlap measure similar to Equation 5.7. The measures vary in the word-to-word similarity used and the weighting scheme. Distributional similarity (such as (Lin, 1998)) and WordNet-based similarity measures (such as (Leacock, Miller, and Chodorow, 1998)) were applied. In addition, the different works vary in

the preprocessing done (tokenization, lemmatization, etc.) and in the corpora used to collect statistics. For this reason it is difficult to compare the performance of the different measure variants of different systems. Nevertheless the reported results were usually comparable with each other, which may suggest that these lexical techniques are somewhat close to exhausting the potential of current state-of-the-art in lexical based systems.

The proposed lexical models are, after all, quite simple and make many obviously wrong assumptions. Clearly, there is an upper bound of performance one would expect from a system working solely at the lexical level (see the analysis in Section 3.5.3 and in (Bar-Haim, Szpektor, and Glickman, 2005)). Incorporating additional linguistic levels into the probabilistic entailment model, such as syntactic matching, co-reference resolution and word sense disambiguation in such a way that they improve system performance becomes a challenging target for future research.

*“Adah and Zillah, hear my voice; ye wives of Lamech,
harken unto my speech”*

— (Genesis 4:23)

*“Tell it not in Gath, publish it not in the streets of Askelon;
lest the daughters of the Philistines rejoice, lest the daughters
of the uncircumcised triumph.”*

— (Samuel 2 1:20)

Chapter 6

Acquiring Lexical Entailment Relations

6.1 Motivation

Performing textual inferences at a broad scale, as needed for applications, requires a large knowledge base of entailment patterns in the language (such as directional and symmetric paraphrase relations). As noted in Chapter 5, acquisition of such patterns and lexical entailment relations is an important building block of textual entailment systems. Furthermore, in (Dagan and Glickman, 2004) we propose a generic framework for modeling textual entailment that recognizes language variability at a shallow

semantic level and relies on a knowledge base of paraphrase patterns. Consequently, acquisition of such paraphrase patterns is of great significance. In this chapter we focus on the acquisition rather than the inference and propose an algorithm for identifying lexical paraphrases within a single corpus, focusing on the extraction of verb paraphrases.

Most previous approaches detect individual paraphrase instances within a pair (or set) of comparable corpora, each of them containing roughly the same information, and rely on the substantial level of correspondence of such corpora (see Section 2.2.4). We present a novel method that successfully detects isolated paraphrase instances within a single corpus without relying on any a-priori structure and information. The goal of our research is to explore the potential of learning paraphrases within a single corpus. Clearly, requiring a pair (or set) of comparable corpora is a disadvantage, since such corpora do not exist for all domains, and are substantially harder to assemble. On the other hand, the approach of detecting actual paraphrase instances seems to have high potential for extracting reliable paraphrase patterns. We therefore developed a method that detects concrete paraphrase *instances* within a single corpus. Such paraphrase instances can be found since a coherent domain corpus is likely to include repeated references to the same concrete facts or events, even though they might be found within generally different stories. For example, Figure 6.1 shows two distinct stories that both refer to a common event in slightly different wordings. Our *instance based* approach identifies such instances and correctly infers, in this example, that split and separated are paraphrases.

As a first attempt in this direction our algorithm was restricted to identify lexical paraphrases of verbs, in order to study whether the approach as a whole is at all feasible. The challenge addressed by our algorithm is to identify isolated paraphrase instances that describe the *same* fact or event within a single corpus. Such paraphrase instances need to be distinguished from instances of *distinct* facts that are described

20-08-1996	16-09-1996
<p>... The broadcast, which gave no source for the information, followed a flurry of rumours that Lien had arrived in various European nations. China regards Nationalist-ruled Taiwan as a rebel province ineligible for foreign ties and has sought to isolate it diplomatically since a civil war separated them in 1949. Adomaitis said Ukraine maintains only economic relations with Taiwan with no political or diplomatic ties ...</p>	<p>“I recognise there are political issues, but I nevertheless see it as a golden opportunity for Taiwan to increase its role in this important international organisation, and to play the part that it should as a major Asian economy,” Summers said. China, which has regarded Taiwan as a rebel province since a civil war split them in 1949, says the island is not entitled to membership as a sovereign nation in international bodies. Beijing has said it would accept Taiwan’s membership in the WTO as a customs territory, but not before China itself is allowed to join the world trade club.</p>

Figure 6.1: Example of extracting the lexical paraphrase ⟨separate, split⟩ from distinct stories

in similar terms. These goals are achieved through a combination of statistical and linguistic filters and a probabilistically motivated paraphrase likelihood measure. We found that the algorithmic computation needed for detecting such local paraphrase instances across a single corpus should be quite different than previous methods developed for comparable corpora, which largely relied on a-priori knowledge about the correspondence between the different stories from which the paraphrase instances are extracted.

We have further compared our method to the vector-based approach of (Lin and Pantel, 2001). The precision of the two methods on common verbs was comparable, but they exhibit some different behaviors. In particular, our instance-based approach seems to help assessing the reliability of candidate paraphrases, which is more difficult to assess by global similarity measures such as the measure of Lin and Pantel.

subject	secretary_general_boutros_boutros_ghali	subject	iraqi_force
object	implementation_of_deal	object	kurdish_rebel
modifier	after	pp-on	august_31
(A) verb:	delay	(B) verb:	attack

Figure 6.2: Extracted verb instances for sentence “But U.N. Secretary-General Boutros Boutros-Ghali delayed implementation of the deal after Iraqi forces attacked Kurdish rebels on August 31.”

6.2 Algorithm

Our proposed algorithm identifies candidates of corresponding verb paraphrases within pairs of sentences. We define a *verb instance pair* as a pair of occurrences of two distinct verbs in the corpus. A *verb type pair* is a pair of verbs detected as a candidate lexical paraphrase.

6.2.1 Preprocessing and Representation

Our algorithm relies on a syntactic parser to identify the syntactic structure of the corpus sentences, and to identify verb instances. We treat the corpus uniformly as a set of distinct sentences, regardless of the document or paragraph they belong to. For each verb instance we extract the various syntactic components that are related directly to the verb in the parse tree. For each such component we extract its lemmatized head, which is possibly extended to capture a semantically specified constituent. We extended the heads with any lexical modifiers that constitute a multi-word term, noun-noun modifiers, numbers and prepositional ‘of’ complements. Verb instances are represented by the vector of syntactic modifiers and their lemmatized fillers. For illustration, Figure 6.2 shows an example sentence and the vector representations for its two verb instances.

6.2.2 Identifying Candidate Verb Instance Pairs (Filtering)

We apply various filters in order to verify that two verb instances are likely to be paraphrases describing the same event. This is an essential part of the algorithm since we do not rely on the high a-priori likelihood for finding paraphrases in matching parts of comparable corpora.

We first limit our scope to pairs of verb instances that share a common (extended) subject and object which are not pronouns. Otherwise, if either the subject or object differ between the two verbs then they are not likely to refer to the same event in a manner that allows substituting one verb with the other. Additionally, we are interested in identifying sentence pairs with a significant overall term overlap, which further increases paraphrase likelihood for the same event. This is achieved with a standard (Information Retrieval style) vector-based approach, with tf-idf term weighting

- $tf(w) = freq(w)$ in sentence
- $idf(w) = \log(N/freq(w))$ in corpus) where N is the total number of tokens in the corpus.

Sentence overlap is measured simply as the dot product of the two vectors. We intentionally disregard any normalization factor (such as in the cosine measure) in order to assess the absolute degree of overlap, while allowing longer sentences to include also non-matching parts that might correspond to complementary aspects of the same event. Verb instance pairs whose sentence overlap is below a specified threshold are filtered out.

An additional assumption is that events have a unique propositional representation and hence verb instances with contradicting vectors are not likely to describe the same event. We therefore filter verb instance pairs with contradicting propositional information - a common syntactic relation with different arguments. As an example,

the sentence “Iraqi forces captured Kurdish rebels on August 29.” Has a contradicting ‘on’ preposition argument with the sentence from Figure 6.2(B) (“August 29” vs. “August 31”).

6.2.3 Computing the Paraphrase Score of Verb Instance Pairs

Given a verb instance pair (after filtering), we want to estimate the likelihood that the two verb instances are paraphrases of the same fact or event. We thus assign a paraphrase likelihood score for a given verb instance pair I_{v_1, v_2} , which corresponds to instances of the verb types v_1 and v_2 with overlapping syntactic components p_1, p_2, \dots, p_n . The score corresponds (inversely) to the estimated probability that such overlap had occurred by chance in the entire corpus, capturing the view that a low overlap probability (i.e., low probability that the overlap is due to chance) correlates with paraphrase likelihood. We estimate the overlap probability by assuming independence of the verb and each of its syntactic components as follows:

$$\begin{aligned} P(I_{v_1, v_2}) &= P(\text{overlap}) = P(v_1, p_1, \dots, p_n) P(v_2, p_1, \dots, p_n) \\ &= P(v_1) P(v_2) \prod_{i=1}^n P(p_i)^2 \end{aligned} \quad (6.1)$$

where the probabilities were calculated using Maximum Likelihood estimates based on the verb and argument frequencies in the corpus.

6.2.4 Computing Paraphrase Score for Verb Type Pairs

When computing the score for a verb type pair we would like to accumulate the evidence from its corresponding verb instance pairs. Following the vein of the previous sub-section we try to estimate the joint probability that these different instance

pairs occurred by chance. Assuming instance independence, we would like to multiply the overlap probabilities obtained for all instances. We have found, though, that verb instance pairs whose two verbs share the same subject and object are far from being independent (there is a higher likelihood to obtain additional instances with the same subject-object combination). To avoid complex modeling of such dependencies we picked only one verb instance pair for each subject-object combination, taking the one with lowest probability (highest score). This yields the set $T(v_1, v_2) = (I_1, \dots, I_n)$ of best scoring (lowest probability) instances for each distinct subject and object components. Assuming independence of occurrence probability of these instances, we estimate the probability $P(T(v_1, v_2)) = \prod P(I_i)$, where $P(I)$ is calculated by Equation (6.1) above. The score of a verb type pair is given by: $\text{score}(v_1, v_2) = -\log P(T(v_1, v_2))$.

6.3 Evaluation and Analysis

6.3.1 Setting

We ran our experiments on the first 15-million word (token) subset of the Reuters Corpus (Rose, Stevenson, and Whitehead, 2002). The corpus sentences were parsed using the Minipar dependency parser (Lin, 1993). 6,120 verb instance pairs passed filtering (with overlap threshold set to 100). These verb instance pairs derive 646 distinct verb type pairs, which were proposed as candidate lexical paraphrases along with their corresponding paraphrase score.

The correctness of the extracted verb type pairs was evaluated over a sample of 215 pairs (one third of the complete set) by two human judges, where each judge evaluated one half of the sample. In a similar vein to related work in this area, judges were instructed to evaluate a verb type pair as a *correct* paraphrase only if the following

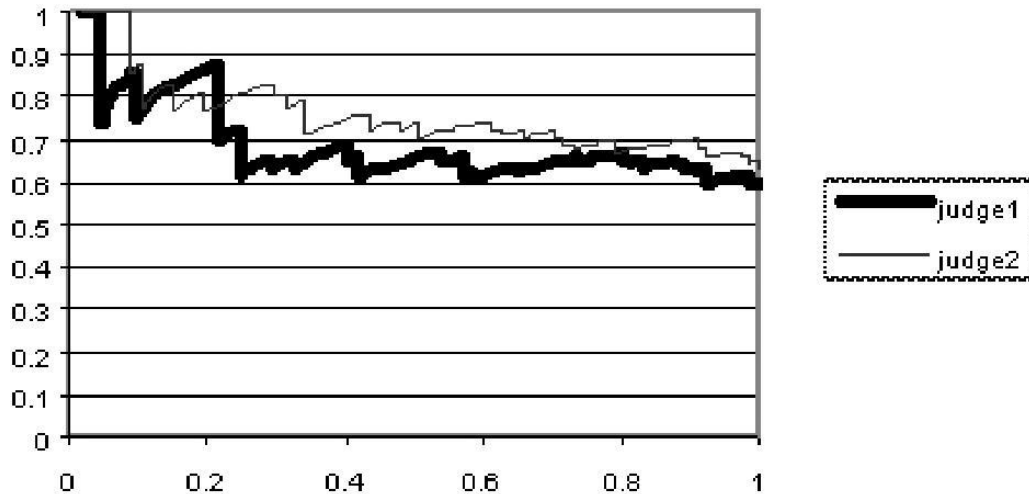


Figure 6.3: Precision (y axis) recall (x axis) curves of system paraphrases by judge (verb type pairs sorted by system score)

condition holds: one of the two verbs can replace the other within some sentences such that the meaning of the resulting sentence will entail the meaning of the original one. To assist the judges in assessing a given verb type pair they were presented with example sentences from the corpus that include some matching contexts for the two verbs (e.g., sentences in which both verbs have the same subject or object). Notice that the judgment criterion allows for directional paraphrases, such as $\langle \text{invade, enter} \rangle$ or $\langle \text{slaughter, kill} \rangle$, where the meaning of one verb entails the meaning of the other, but not vice versa.

6.3.2 Results of the Paraphrase Identification Algorithm

Figure 6.3 shows the precision vs. recall results for each judge over the given test-sets. The evaluation was conducted separately also by the author on the full set of 646 verb pairs, obtaining comparable results to the independent evaluators. In terms of agreement, the Kappa value (measuring pair wise agreement discounting

1	⟨fall, rise⟩	-	32	⟨accuse, blame⟩	+
2	⟨close, end⟩	+	62	⟨honor, honour⟩	+
3	⟨post, report⟩	+	92	⟨raise, set⟩	-
4	⟨recognise, recognize⟩	+	122	⟨advance, rise⟩	+
5	⟨fire, launch⟩	+	152	⟨belt, sandwich⟩	-
6	⟨drop, fall⟩	+	182	⟨benefit, bolster⟩	+
7	⟨regard, view⟩	+	212	⟨boost, propel⟩	+
8	⟨cut, lower⟩	+	242	⟨approve, authorize⟩	+
9	⟨rise, shed⟩	-	272	⟨fall, slip_in⟩	+
10	⟨fall, slip⟩	+	302	⟨kill, slaughter⟩	+
11	⟨edge, rise⟩	+	332	⟨inform, notify⟩	+
12	⟨increase, rise⟩	+	362	⟨bring, take⟩	+
13	⟨gain, rise⟩	+	392	⟨raise, spark⟩	+
14	⟨ease, rise⟩	-	422	⟨note, say⟩	+
15	⟨fall, shed⟩	+	452	⟨develop, have⟩	-
16	⟨fall, lose⟩	+	482	⟨export, load⟩	-
17	⟨add, rise⟩	+	512	⟨capture, return⟩	-
18	⟨end, end_up⟩	+	542	⟨downgrade, relax⟩	+
19	⟨narrow, widen⟩	-	572	⟨attack, leave⟩	-
20	⟨close, rise⟩	+	602	⟨create, establish⟩	+
21	⟨tell, warn⟩	+	632	⟨announce, give⟩	-

Table 6.1: Example of system output with judgments

chance occurrences) between the author and the independent evaluators’ judgments were 0.61 and 0.63, which correspond to a substantial agreement level (Landis and Koch, 1997). The overall precision for the complete test sample is 61.4% accuracy, with a confidence interval of [56.1,66.7] at the 0.05 significance level. Table 6.1 shows the top 10 lexical paraphrases, and a sample of the remaining ones, achieved by our system along with the annotators’ judgments. Table 6.2 shows correct sentence pairs describing a common event, which were identified by our system as candidate paraphrase instances.

An analysis of the incorrect paraphrases showed that roughly one third of the errors captured verbs with contradicting semantics or antonyms (e.g., ⟨rise, fall⟩, ⟨buy, sell⟩,

correct paraphrase instance pairs	
Campbell is buying Erasco from Grand Metropolitan Plc of Britain for about \$210 million.	Campbell is purchasing Erasco from Grand Metropolitan for approximately US\$210 million.
The stock of Kellogg Co. dropped Thursday after the giant cereal maker warned that its earnings for the third quarter will be 20 percent below a year ago.	The stock of Kellogg Co. fell Thursday after it warned about lower earnings this year . . .
Ieng Sary on Wednesday formally announced his split with top Khmer Rouge leader Pol Pot, and said he had formed a rival group called the Democratic National United Movement.	In his Wednesday announcement Ieng Sary, who was sentenced to death in absentie for his role in the Khmer Rouge's bloody rule, confirmed his split with paramount leader Pol Pot.
Slovenian President Milan Kucan opened a second round of consultations with political parties on Monday to try to agree on a date for a general election which must take place between October 27 and December 8.	- Slovenian President Milan Kucan on Monday started a second round of consultations with political parties concerning the election date.
misleading instance pairs	
Last Friday, the United States announced punitive charges against China's 1996 textile and apparel quotas . . .	China on Saturday urged the United States to rescind punitive charges against Beijing's 1996 textile and apparel quotas . . .
Municipal bond yields dropped as much as 15 basis points in the week ended Thursday, erasing increases from the week before.	Municipal bond yields jumped as much as 15 basis points over the week ended Thursday . . .
Rand Financials notably bought October late while Chicago Corp and locals lifted December into by stops.	Rand Financials notably sold October late while locals pressured December.
French shares opened lower, ignoring gains on Wall Street and other European markets, due to renewed pressure on the franc and growing worries about possible strike action in the autumn, dealers said.	French shares closed sharply lower on Wednesday due to a weaker franc amid evaporating hopes of a German rate cut on Thursday, but the market managed to remain above the 2,000 level and did keep some of Tuesday's gains.

Table 6.2: Examples of instance pairs

⟨capture, evacuate⟩) and another third were verbs that tend to represent correlated events with strong semantic similarity (e.g., ⟨warn, attack⟩, ⟨reject, criticize⟩). These cases are indeed quite difficult to distinguish from true paraphrases since they tend to occur in a corpus with similar overlapping syntactic components and within quite similar sentences. Table 6.2 also shows examples of misleading sentence pairs demonstrating the difficulties posed by such instances. As our goal is to extract generic paraphrase patterns our evaluation was performed at the verb type level. We have not evaluated directly the correctness of the individual paraphrase instance pairs extracted by our method (i.e., whether the two instances in a paraphrase pair indeed refer to the same fact). Finally, a general problematic (and rarely addressed) issue in this area of research is how to evaluate the coverage or recall of the extraction method relative to a given corpus.

6.3.3 Comparison with (Lin & Pantel 2001)

We implemented the algorithm of (Lin & Pantel 2001), denoted here as the LP algorithm, and computed their similarity score for each pair of verb types in the corpus (see also Section 2.2.4). To implement the method for lexical verb paraphrases, each verb type was considered as a distinct path whose subject and object play the roles of the X and Y slots.

As it turned out, the similarity score of LP does not behave uniformly across all verbs. For example, many of the top 20 highest scoring verb pairs are quite erroneous (see Figure 6.3), and do not constitute lexical paraphrases (compare with the top scoring verb pairs for our system in Figure 6.1). The similarity scores do seem meaningful within the context of a single verb v , such that when sorting all other verbs by the LP score of their similarity to v correct paraphrases are more likely to occur in the upper part of the list. Yet, we are not aware of a criterion that predicts whether a certain verb has few good paraphrases, many or none. Given this behavior

⟨misread, misjudge⟩	0.629	⟨flatten, steepen⟩	0.207
⟨barricade, sandbag⟩	0.290	⟨mainline, pip⟩	0.205
⟨disgust, mystify⟩	0.278	⟨misinterpret, relieve⟩	0.202
⟨jack, decontrol⟩	0.274	⟨remarry, flaunt⟩	0.192
⟨pollinate, pod⟩	0.256	⟨distance, dissociate⟩	0.187
⟨mark_down, decontrol⟩	0.238	⟨trumpet, drive_home⟩	0.180
⟨subsidize, subsidise⟩	0.223	⟨marshal, beleaguer⟩	0.173
⟨wake_up, divine⟩	0.218	⟨dwell_on, feed_on⟩	0.172
⟨thrill, personify⟩	0.212	⟨scrutinize, misinterpret⟩	0.164
⟨mark_up, decontrol⟩	0.208	⟨disable, counsel⟩	0.164

Table 6.3: Top 20 verb pairs from similarity system along with their similarity score

of the LP score we created a test sample for the LP algorithm by randomly selecting verb pairs of equivalent similarity rankings relative to the original test sample. Notice that this procedure is favorable to the LP method for it is evaluated at points (verb and rank) that were predicted by our method to correspond to a likely paraphrase.

The resulting 215 verb pairs were evaluated by the judges along with the sample for our method, while the judges did not know which system generated each pair. The overall precision on the LP method for the sample was 51.6%, with a confidence interval of [46.1,57.1] at the 0.05 significance level. The LP results for this sample were thus about 10 points lower than the results for our comparable sample, but the two statistical confidence intervals overlap slightly. It is interesting to note that the precision of the LP algorithm over all pairs of rank 1 was also 51%, demonstrating that just rank on its own is not a good basis for paraphrase likelihood. Figure 6.4 shows overall recall vs. precision from both judges for the two systems. The results above show that the precision of the vector-based LP method may be regarded as comparable to our instance-based method, in cases where one of the two verbs was identified by our method to have a corresponding number of paraphrases. The obtained level of accuracy for these cases is substantially higher than for the top scoring pairs by LP.

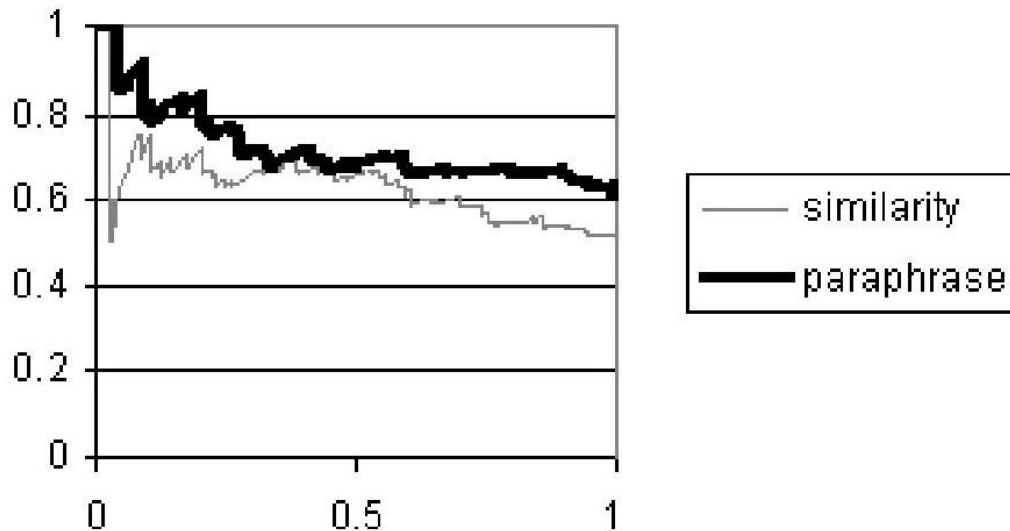


Figure 6.4: Precision recall curve for our paraphrase method and LP similarity

This suggests that our approach can be combined with the vector-based approach and thus obtain higher reliability for verb pairs that were extracted from actual paraphrase instances.

6.4 Conclusions

This work presents an algorithm for extracting lexical verb paraphrases from a single corpus. To the best of our knowledge, this is the first attempt to identify actual paraphrase instances in a single corpus and to extract paraphrase patterns directly from them. The evaluation suggests that such an approach is indeed viable, based on algorithms that are geared to overcome many of the misleading cases that are typical for a single corpus (in comparison to comparable corpora). Furthermore, a preliminary comparison suggests that verb pairs extracted by our instance-based approach are more reliable than those based on distributional vector similarity. As a result, an instance-based approach may be combined with a vector-based approach in

order to assess better the paraphrase likelihood for many verb pairs. Future research is needed to extend the approach to handle more complex paraphrase structures and to increase its performance by relying on additional sources of evidence.

Sir Humphrey: “The identity of the official whose alleged responsibility for this hypothetical oversight has been the subject of recent discussion, is NOT shrouded in quite such impenetrable obscurity as certain previous disclosures may have led you to assume, but not to put too fine a point on it, the individual in question is, it may surprise you to learn, one whom your present interlocutor is in the habit of identifying by means of the perpendicular pronoun.”

James Hacker: “I beg your pardon?”

Sir Humphrey: “It was ...I.”

— from the BBC comedy series *Yes Minister*

Chapter 7

Conclusions

7.1 The Textual Entailment Task and Evaluation Framework

This dissertation introduced, in Chapter 3, the notion of textual entailment as a generic empirical task that captures major semantic inferences across applications. The PASCAL *Recognising Textual Entailment* (RTE) Challenge was an initial attempt to form a common benchmark dataset for textual entailment evaluation. The

high level of interest in the challenge, demonstrated by the submissions from 17 diverse groups and noticeable interest in the research community, suggest that textual entailment indeed captures highly relevant core semantic tasks.

The results obtained by the participating systems may be viewed as typical for a new and relatively difficult task (cf. for example the history of the Message Understanding Conferences (MUC) benchmarks (Sundheim and Chinchor, 1993)). Overall performance figures for the better systems were significantly higher than some baselines. Yet, the absolute numbers are relatively low, with small, though significant, differences between systems. Interestingly, system complexity and sophistication of inference did not correlate fully with performance, where some of the best results were obtained by rather naïve lexically-based systems. The fact that quite sophisticated inference levels were applied by some groups, with 6 systems applying logical inference, provides an additional indication that applied NLP research is progressing towards deeper semantic reasoning. Additional refinements are needed though to obtain sufficient robustness for the Challenge types of data. Further detailed analysis of systems performance, relative to different types of examples and entailment phenomena, are likely to yield future improvements.

The introduction of the textual entailment task and the RTE dataset has already raised considerable interest and impact on the NLP community. We first presented the notion of textual entailment in a workshop paper in January 2004 (Dagan and Glickman, 2004) and in June 2004 we first announced the RTE challenge and made available a portion of the dataset. The challenge raised noticeable attention in the research community. It was followed by an ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment¹. As well as over a dozen papers specifically on the topic of textual entailment appeared in this year's main conferences (Bos and

¹<http://acl.ldc.upenn.edu/W/W05/#W05-1200>

Markert, 2005; de Salvo Braz et al., 2005a; de Salvo Braz et al., 2005b; Geffet and Dagan, 2005; Glickman, Dagan, and Koppel, 2005a; Haghighi, Ng, and Manning, 2005; Kouylekov and Magnini, 2005; Raina, Ng, and Manning, 2005; Rus, Graesserand, and Desai, 2005). Furthermore, Recent calls for papers list textual entailment as one of their topics – RANLP-05, EACL-06, ICos-5 (2006) and NL-KR 06. All these provide a clear indication for the quick establishment of this novel research subject.

7.2 The Probabilistic Setting and Derived Models

In Chapter 4 we proposed a general probabilistic setting that formalizes the notion of textual entailment. And in Chapter 5 we described two lexical models derived from the probabilistic setting which utilize co-occurrence statistics. Although our proposed models are relatively simple, as they do not rely on syntactic or other deeper analysis, they nevertheless achieved competitive results on the PASCAL RTE challenge. These results may suggest that the proposed probabilistic framework is a promising basis for improved implementations that would incorporate deeper types of information. It would also be interesting for future work to possibly combine the two models and thus capture entailment from a specific text term as well as the impact of the entire context within the given text.

One can make an analogy here to the history of statistical machine translation. The introduction of the statistical (noisy channel) setting for MT in (Brown et al., 1990) provides the necessary grounding for probabilistic modeling of machine translation and the development of concrete models. The first IBM models 1 through 5 (Brown et al., 1993) were based on lexical substitutions and permutations and did not involve any syntactic analysis or information. These and later lexical models were among the top performing systems for almost a decade. At that point, word-aligned corpora have been effectively exploited for constructing phrase substitution models

which have significantly outperformed word-based models (Och and Ney, 2004). Only recently, models taking a step toward linguistic syntax proved successful (Chiang, 2005). Or quoting from (Knight and Marcu, 2004) – “Research in word and phrase based decoding is in its teens; research in syntax-based decoding is in its infancy.” As modeling textual entailment is still in its infancy, we expect to see a similar course of events. Statistical machine translation models have already been applied successfully in a monolingual setting to model language variability (e.g (Quirk, Brockett, and Dolan, 2004; Bayer et al., 2005)) (see also Chapter 2). However we believe that semantic inference is more than just monolingual translation (semantic equivalence) and hence modeling textual entailment might require more sophisticated models than individually translating phrases and reordering them. Furthermore, it is not clear if there would be available large training data for textual entailment as of the scale available in MT. As a consequence, it is possible that the creation of this new discipline will leave researchers no much of a choice but to work on the core semantic issues of applied natural language and trigger ‘deeper’ linguistic modeling and techniques.

7.3 Learning Entailment Rules

In Chapter 6 we described a novel algorithm for learning lexical entailment relations within a single corpus. The instance based approach identifies actual paraphrase instances that describe the same fact or event rather than comparing typical contexts in a global manner and thus addresses some of the drawbacks of distributional similarity. As a direct carry-over of our work, the ideas and findings of our work sparked the fruitful collaborative effort of TEASE (Szpektor et al., 2004) – a scalable unsupervised web-based method to learn syntactic paraphrase patterns (see also Chapter 2).

References

- [Allen1995] Allen, James. 1995. *Natural language understanding (2nd ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.
- [Bacchus1990] Bacchus, Fahiem. 1990. *Representing and reasoning with probabilistic knowledge: a logical approach to probabilities*. MIT Press, Cambridge, MA, USA.
- [Bar-Haim, Szpektor, and Glickman2005] Bar-Haim, Roy, Idan Szpektor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 55–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- [Barzilay and Lee2003] Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL 2003: Main Proceedings*, pages 16–23.
- [Barzilay and McKeown2001] Barzilay, Regina and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *ACL*, pages 50–57.
- [Barzilay, McKeown, and Elhadad1999] Barzilay, Regina, Kathleen McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *ACL*.
- [Bayer et al.2005] Bayer, Samuel, John Burger, Lisa Ferro, John Henderson, and Alexander Yeh. 2005. Mitre’s submissions to the eu pascal rte challenge. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).
- [Berger and Lafferty1999] Berger, Adam and John Lafferty. 1999. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229, New York, NY, USA. ACM Press.
- [Blackburn and Bos2005] Blackburn, Patrick and Johan Bos. 2005. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI.
- [Bos and Markert2005] Bos, Johan and Katja Markert. 2005. Recognising textual entailment with logical inference techniques. In *EMNLP*.

- [Brown et al.1990] Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.
- [Brown et al.1993] Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.
- [Califf and Mooney2003] Califf, Mary Elaine and Raymond J. Mooney. 2003. Bottom-up relational learning of pattern matching rules for information extraction. *J. Mach. Learn. Res.*, 4:177–210.
- [Chai and Bierman1997] Chai, Joyce Yue and Alan W. Bierman. 1997. The use of lexical semantics in information extraction. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors, *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Association for Computational Linguistics, New Brunswick, New Jersey, pages 61–70.
- [Chiang2005] Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- [Chierchia and McConnell-Ginet2000] Chierchia, Gennaro and Sally McConnell-Ginet. 2000. *Meaning and grammar (2nd ed.): an introduction to semantics*. MIT Press, Cambridge, MA, USA.
- [Church1988] Church, Kenneth Ward. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing*, pages 136–143, Morristown, NJ, USA. Association for Computational Linguistics.
- [Collins1997] Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.
- [Contact2003] Contact. 2003. *Newsletter of the Association of Teachers of English as a Second Language of Ontario*, 29(3):11–12.

- [Corley and Mihalcea2005] Corley, Courtney and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- [Crouch et al.2003] Crouch, Dick, Cleo Condoravdi, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. HLT-NAACL Workshop on Text Meaning.
- [Daelemans, Höthker, and Tjong Kim Sang2004] Daelemans, Walter, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.
- [Dagan and Glickman2004] Dagan, Ido and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. PASCAL workshop on Text Understanding and Mining.
- [Dagan, Glickman, and Magnini2005] Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).
- [Dalmas and Webber2005] Dalmas, Tiphaine and Bonnie Webber. 2005. Using information fusion for open domain question answering. In *Proceedings of the KRAQ workshop, IJCAI*, pages 81–88.
- [de Salvo Braz et al.2005a] de Salvo Braz, Rodrigo, Roxana Girju, Vasin Punyakanok, Dan Roth, and Mark Sammons. 2005a. An inference model for semantic entailment in natural language. In *AAAI*, pages 1043–1049.
- [de Salvo Braz et al.2005b] de Salvo Braz, Rodrigo, Roxana Girju, Vasin Punyakanok, Dan Roth, and Mark Sammons. 2005b. Knowledge representation for semantic entailment and question-answering. In *IJCAI-05 Workshop on Knowledge and Reasoning for Answering Questions*, pages 71–80.
- [Forman2003] Forman, George. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305.
- [Frege1892] Frege, Gottlob. 1892. On sense and reference. Reprinted in P. Geach and M. Black, eds., *Translations from the Philosophical Writings of Gottlob Frege*. 1960.

- [Gaizauskas and Wilks1998] Gaizauskas, Robert and Yorick Wilks. 1998. Information extraction: Beyond document retrieval. *Journal of Documentation*, 54(1):70–105.
- [Geffet and Dagan2004] Geffet, Maayan and Ido Dagan. 2004. Feature vector quality and distributional similarity. In *Proceedings of Coling 2004*, pages 247–253, Geneva, Switzerland, Aug 23–Aug 27. COLING.
- [Geffet and Dagan2005] Geffet, Maayan and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- [Glickman and Dagan2004] Glickman, Oren and Ido Dagan, 2004. *Recent Advances in Natural Language Processing III*, chapter Acquiring lexical paraphrases from a single corpus, pages 81–90. John Benjamins.
- [Glickman, Dagan, and Koppel2005a] Glickman, Oren, Ido Dagan, and Moshe Koppel. 2005a. A probabilistic classification approach for lexical textual entailment. In *AAAI*, pages 1050–1055.
- [Glickman, Dagan, and Koppel2005b] Glickman, Oren, Ido Dagan, and Moshe Koppel. 2005b. Web based probabilistic textual entailment. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).
- [Haghighi, Ng, and Manning2005] Haghighi, Aria D., Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via graph matching. In *EMNLP*.
- [Halpern1990] Halpern, Joseph Y. 1990. An analysis of first-order logics of probability. *Artif. Intell.*, 46(3):311–350.
- [Harabagiu et al.2000] Harabagiu, Sanda M., Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. Falcon: Boosting knowledge for answer engines. In *TREC*.
- [Harris1968] Harris, Zelig. 1968. *Mathematical Structures of Language*. New York: Wiley.
- [Hirschman and Gaizauskas2001] Hirschman, L. and R. Gaizauskas. 2001. Natural language question answering: the view from here. *Nat. Lang. Eng.*, 7(4):275–300.
- [Hirschman et al.1999] Hirschman, Lynette, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: a reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on*

- Computational Linguistics*, pages 325–332, Morristown, NJ, USA. Association for Computational Linguistics.
- [Hobbs et al.1988] Hobbs, J. R., M. Stickel, P. Martin, and D. Edwards. 1988. Interpretation as abduction. In *Proc. of the 26th ACL*, pages 95–103.
- [Hovy, Hermjakob, and Lin2001] Hovy, Eduard H., Ulf Hermjakob, and Chin-Yew Lin. 2001. The use of external knowledge of factoid QA. In *Text REtrieval Conference*.
- [Jijkoun and de Rijke2005] Jijkoun, Valentin and Maarten de Rijke. 2005. Recognizing textual entailment using lexical similarity. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).
- [Joachims2005] Joachims, Thorsten. 2005. A support vector method for multivariate performance measures. In *ICML*.
- [Knight and Marcu2004] Knight, Keven and Daniel Marcu. 2004. Machine translation in the year 2004. In *ICASSP*.
- [Knight and Marcu2002] Knight, Kevin and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107.
- [Kouylekov and Magnini2005] Kouylekov, Milen and Bernardo Magnini. 2005. Tree edit distance for textual entailment. In *Recent Advances in Natural Language Processing (RANLP)*.
- [Kwok, Etzioni, and Weld2001] Kwok, Cody C. T., Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. In *World Wide Web*, pages 150–161.
- [Landis and Koch1997] Landis, J. R. and G. G. Koch. 1997. The measurements of observer agreement for categorical data. *Biometrics*, 33:159–174.
- [Leacock, Miller, and Chodorow1998] Leacock, Claudia, George A. Miller, and Martin Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.*, 24(1):147–165.
- [Lin2004] Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

- [Lin1993] Lin, Dekang. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 112–120, Morristown, NJ, USA. Association for Computational Linguistics.
- [Lin1998] Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.
- [Lin and Pantel2001] Lin, Dekang and Patrik Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 4(7):343–360.
- [Lukasiewicz1970] Lukasiewicz, Jan. 1970. *Selected Works*. North Holland, London.
- [Marsi and Krahmer2005] Marsi, Erwin and Emiel Krahmer. 2005. Classification of semantic relations by humans and machines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 1–6, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- [Matthew W. Bilotti and Lin2004] Matthew W. Bilotti, Boris Katz and Jimmy Lin. 2004. What works better for question answering: Stemming or morphological query expansion. SIGIR 2004 Workshop on Information Retrieval for Question Answering.
- [McCallum and Nigam1998] McCallum, Andrew and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *Proc. AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48.
- [Miller1995] Miller, G. A. 1995. WordNet: A Lexical Databases for English. *Communications of the ACM*, pages 39–41, November.
- [Moldovan and Rus2001] Moldovan, Dan I. and Vasile Rus. 2001. Logic form transformation of wordnet and its applicability to question answering. In *ACL*, pages 394–401.
- [Monz and de Rijke2001] Monz, Christof and Maarten de Rijke. 2001. Light-weight entailment checking for computational semantics. In P. Blackburn and M. Kohlhase, editor, *Proceedings ICoS-3*.
- [Negri2004] Negri, Matteo. 2004. Sense-based blind relevance feedback for question answering. SIGIR 2004 Workshop on Information Retrieval for Question Answering.
- [Nie and Brisebois1996] Nie, Jian-Yun and Martin Brisebois. 1996. An inferential approach to information retrieval and its implementation using a manual thesaurus. *Artif. Intell. Rev.*, 10(5-6):409–439.

- [Nigam et al.2000] Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134.
- [Och and Ney2003] Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- [Och and Ney2004] Och, Franz Josef and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 4(30):417–449.
- [Pang, Knight, and Marcu2003] Pang, Bo, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *HLT-NAACL*.
- [Papineni et al.2001] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.
- [Ponte and Croft1998] Ponte, Jay M. and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA. ACM Press.
- [Qiu and Frei1993] Qiu, Yonggang and Hans-Peter Frei. 1993. Concept based query expansion. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 160–169. ACM.
- [Quirk, Brockett, and Dolan2004] Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain, July. Association for Computational Linguistics.
- [Radev2000] Radev, Dragomir. 2000. A common theory of information fusion from multiple text sources. In Laila Dybkjaer, Koiti Hasida, and David Traum, editors, *Proceedings of the First SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, Somerset, New Jersey, pages 74–83.

- [Raina, Ng, and Manning2005] Raina, Rajat, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via learning and abductive reasoning. In *AAAI*, pages 1099–1105.
- [Reynar and Ratnaparkhi1997] Reynar, Jeffrey C. and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*, pages 16–19, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Robertson and Jones1988] Robertson, Stephen E. and Karen Sparck Jones. 1988. Relevance weighting of search terms. pages 143–160.
- [Rose, Stevenson, and Whitehead2002] Rose, Tony G., Mark Stevenson, and Miles Whitehead. 2002. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 29–31.
- [Rus, Graesserand, and Desai2005] Rus, Vasile, Art Graesserand, and Kirtan Desai. 2005. Lexico-syntactic subsumption for textual entailment. In *Recent Advances in Natural Language Processing (RANLP)*.
- [Russell1919] Russell, Bertrand. 1919. Descriptions. In *Introduction to Mathematical Philosophy*. George Allen and Unwin, London.
- [Russell and Norvig1995] Russell, Stuart and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ.
- [Saggion et al.2004] Saggion, Horacio, Robert Gaizauskas, Mark Hepple, Ian Roberts, and Mark A. Greenwood. 2004. Exploring the performance of boolean retrieval strategies for open domain question answering. SIGIR Workshop on Information Retrieval for Question Answering.
- [Shinyama et al.2002] Shinyama, Yusuke, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *HLT*, pages 51–58.
- [Sundheim and Chinchor1993] Sundheim, Beth M. and Nancy A. Chinchor. 1993. Survey of the message understanding conferences. In *Proceedings of the DARPA Spoken and Written Language Workshop*.
- [Szpektor et al.2004] Szpektor, Idan, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 41–48, Barcelona, Spain, July. Association for Computational Linguistics.

- [Van Rijsbergen1979] Van Rijsbergen, C. J. 1979. *Information Retrieval, 2nd edition*. Butterworth.
- [Vanderwende, Coughlin, and Dolan2005] Vanderwende, Lucy, Deborah Coughlin, and Bill Dolan. 2005. What syntax can contribute in entailment task. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (and forthcoming LNAI book chapter).
- [Voorhees1994] Voorhees, Ellen M. 1994. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.
- [Voorhees and Harman1999] Voorhees, Ellen M. and Donna Harman. 1999. Overview of the seventh text retrieval conference. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication.
- [Voorhees and Tice2000] Voorhees, Ellen M. and Dawn M. Tice. 2000. Building a question answering test collection. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207, New York, NY, USA. ACM Press.
- [Yang and Chua2002] Yang, Hui and Tat-Seng Chua. 2002. The integration of lexical knowledge and external resources for question answering. In *TREC*.
- [Yang and Pedersen1997] Yang, Yiming and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Zadeh1965] Zadeh, L.A. 1965. Fuzzy Sets. *Information and Control*, 3(8):338–353.