ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n°432:
SMI - Sciences des Métiers de l'Ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité «Informatique temps réel, robotique et automatique - Fontainebleau»

présentée et soutenue publiquement par

Sebastião PAIS

le 06 décembre 2013

Asymmetric Distributional Similarity Measures to Recognize Textual Entailment by Generality

Directeurs de thése: **Robert MAHL et Gaël Harry DIAS**
Co-encadrement de la thèse: **Katarzyna WEGRZYN-WOLSKA**

Jury

Witold KOSINSKI, Professeur, Polish-Japanese Institute of Information Technology	Président & Rapporteur
Robert MAHL, Professeur, CRI, Mines ParisTech	Directeur de la Thèse
Gaël DIAS, Professeur, Département d' Informatique, Univ. of Caen Basse-Normandie	Directeur de la Thèse
Katarzyna WOLSKA-WEGRZYN, Professeur, ESIGETEL	Co-encadrant de la Thèse
Antoine DOUCET, Professeur, Département d' Informatique, Univ. of Caen Basse-Normandie	Rapporteur
João CORDEIRO, Professeur, Département d'Informatique, Univ. of Beira Interior	Examinateur
Zornitsa KOZAREVA, Professeur, Information Sciences Institute, Univ. of Southern California	Examinateur

T H È S E

MINES ParisTech
Centre de Recherche en Informatique (CRI)
35, rue Saint Honoré, 77305 Fontainebleau Cedex, France

If I could reach this stage of my life, this PhD, the writing of this thesis, it is owing to the force that you give me, and to the new meaning that you give to my life, my wife and my daughter. To you I dedicate this thesis. You are responsible for this work, this new victory. I am forever grateful.

Acknowledgements

"Each person who goes in our life goes alone;
it is because each person is unique and no one replacing the other.
Each person who goes in our life passes us alone and not only because each person
leaves a bit of himself and takes a bit of us. (…)"
Charles Chaplin

At all stages of our lives we cross paths with people who contribute directly or indirectly to our success, people who by their importance, became forever in the history of our lives, in our history. Though only my name appears on the cover of this dissertation, many people have contributed to and helped me to bring it up and in its production. I owe my sincere gratitude to all those people who have made this thesis possible and because of whom such experience has been one that I will cherish forever.

I have been fortunate to have an adviser who gave me the freedom, courage and confidence to explore this path on my own, and at the same time the helpful guidance to recover when my steps failed. During all these years of working together, Gaël Dias taught me how to be precise, organized and focused. His support even during the last stages of my thesis and his management and coordination skills have been valuable lessons that I'm sincerely grateful for. *"Muito Obrigado Por Tudo AMIGÃO"*

I am grateful to Robert Mahl and Katarzyna Wegrzyn-Wolska, my adviser and co-advisers, respectively, in the Centre de Recherches en Informatique of Mines Paris (Paristech group of Engineering schools) in Fontainebleau[1]. I am thankful to them for giving me the opportunity to attend Mines ParisTech and get my PhD, one unforgettable experience of great value to me.

My deepest gratitude is to my friend and colleague, Rumen Moraliyski, who has been always there to listen, share ideas and give advices. Through long discussions, critical corrections, creative ideas on technical writings and presentations with him, I have learned priceless lessons during my PhD. I also owe him a debt of gratitude for carefully reading, commenting and revising my writings and slides in all stages of my work.

Thanks to all the members of HULTIG[2], a special thank to João Paulo Cordeiro and Isabel Marcelino.

[1] http://cri.mines-paristech.fr/
[2] http://hultig.di.ubi.pt/

Most importantly, none of this would have been possible without the love and patience of my family. My family to whom this dissertation is dedicated, has been a constant source of love, concern, support and strength for all these years. I would like to express my heart-felt gratitude to my wife and daughter - Carla Pereira and Lara Pais - who have helped and encouraged me. I also offer my deepest thanks and love to my mother - Maria Odete Pais - whose love and support has sustained me during these years. I cannot forget my father and my grand-mother, who already left us, but I know both are looking to me, there, where they are…

I thank you all!
Sebastião Pais

Extended Abstract

"If you can't explain it simply,
you don't understand it well enough."
Albert Einstein

Textual Entaiment (TE) aims to capture major semantic inference needs across applications in Natural Language Processing (NLP). Since 2005, in the TE recognition (RTE) task, systems are asked to automatically judge whether the meaning of a portion of text, the Text - T, entails the meaning of another text, the Hypothesis - H. A number of novel approaches, and improvements in TE technologies demonstrated in recent Recognizing Textual Entaiment (RTE) Challenges are signaling of renewed interest towards a deeper and better understanding of the core phenomena involved in TE.
In line with this direction, in this thesis we focus on a particular case of entailment, entailment by generality. For us, there are various types of implication, range of different levels of entailment reasoning, based on lexical, syntactic, logical and world knowledge at different levels of difficulty. We introduce the paradigm of TE by Generality, which can be defined as the entailment from a specific sentence towards a more general sentence. In this context, the Text T entails the Hypothesis H, because H is more general than T.
We propose an unsupervised and language-independent method to recognize TE by Generality given a case of $Text - Hypothesis$ or $T - H$ where entailment relation holds. To this end, we introduce an Informative Asymmetric Measure (IAM) called Simplified Asymmetric InfoSimba, which we combine with different Asymmetric Association Measures (AAM).
To evaluate the performance of our proposal, we did three experiments:

1. we tested our methodology on all pairs $Text - Hypothesis$ of the Test Set of the first five RTE Challenges;

2. we tested in pairs $Text- > Hypothesis$, where we know the entailment between $Text$ and $Hypothesis$ is by generality;

3. finally, we tested our methodology on 100 pairs $T- > H$ which were randomly extracted from set of pairs submitted in CrowdFlower (60 pairs $T- > H$ Entailment

by Generality and 40 pairs $T->H$ Entailment, but no Generality) and translated into Portuguese by *Google Translate*[1].

To do the experiment with pairs $T->H$, where we know the entailment between T and H is for generality, it was necessary us create a corpus of pairs $Text->Hypothesis$ with Entailment by Generality. This corpus was annotated using the CrowdFlower [2] system, a cheap and fast way to collect annotations from a broad base of paid non-expert contributors over the Web. The corpus is composed of pairs of $Text - Hypothesis$, collected for RTE-1 through RTE-5 challenges. Only positive pairs of TE were submitted to CrowdFlower for annotation, together with a small set of carefully selected cases of known categorization that are used to train the participating annotators and to exercise quality control.

In this work we hypothesize the existence of a special mode of TE, namely Textual Entailment by Generality. Thus, the main contribution of our study is to highlight the importance of this inference mechanism. Consequently, the new annotation data seems to be a valuable recourse for the community.

Keywords: Natural Language Processing, Textual Entailment, Recognizing Textual Entailment by Generality, Word Similarity, Informative Asymmetric Measure, Asymmetric Association Measure

[1] `https://translate.google.pt/` [Last access: 21th December, 2013]

[2] CrowdFlower service provides a crowdsourcing interface to Amazon Mechanical Turk (MTurk) for non-US citizens - `http://crowdflower.com/` [Last access: 14th December, 2013]

Contents

Contents ix

List of Figures xiii

List of Tables xv

List of Abbreviations xvii

1 Introduction 1

1.1 Natural Language Processing . . . 1

1.1.1 Historical Review . . . 2

1.1.2 Applications of Natural Language Processing . . . 4

1.2 Motivation and Rationale . . . 4

1.3 Linguistic Notion of Entailment . . . 8

1.3.1 Variants of the Entailment . . . 8

1.4 Context of Textual Entailment . . . 10

1.4.1 Probabilistic Textual Entailment . . . 11

1.4.2 Recognizing Textual Entailment . . . 12

1.5 Our Proposal for RTE by Generality . . . 13

1.6 Structure of the Thesis . . . 14

2 Related Work 17

2.1 Overview of the First Five RTE Challenges . . . 17

2.1.1 Datasets and Annotations . . . 18

2.1.1.1 RTE-1 . . . 18

2.1.1.2 RTE-2 . . . 21

2.1.1.3 RTE-3 . . . 25

2.1.1.4 RTE-4 . . . 27

2.1.1.5 RTE-5 . . . 29

2.1.1.6 Summary . . . 31

2.1.2 Relevant Resources and Tools . . . 33

2.1.2.1 Evaluation Measures . . . 33
2.1.2.2 First Challenge . . . 33
2.1.2.3 Second Challenge . . . 35
2.1.2.4 Third Challenge . . . 36
2.1.2.5 Fourth Challenge . . . 38
2.1.2.6 Fifth Challenge . . . 40
2.1.2.7 Summary . . . 41
2.2 Unsupervised Language-Independent Methodologies for RTE . . . 42

3 Corpus construction 45
3.1 Crowdsourcing . . . 45
3.2 Quality Control of Crowdsourced Data . . . 46
3.3 Building Methodology . . . 47
3.4 Quantitative Analysis . . . 49

4 Our Methodology for RTE by Generality 51
4.1 Contextual Word Similarity . . . 51
4.1.1 Applications of Word Similarity . . . 52
4.1.2 Co-occurrence relations . . . 53
4.1.2.1 Non-grammatical relations . . . 53
4.1.3 Asymmetric Word Similarities . . . 55
4.1.3.1 Asymmetric Association Measures . . . 55
4.1.3.2 Asymmetric Attributional Word Similarities . . . 57
4.2 Asymmetry between Words . . . 60
4.3 Asymmetry between Sentences . . . 61
4.4 Three Levels of Pre-Processing . . . 62
4.4.1 Multiword Units Identification . . . 63
4.5 Sample of Calculation for Identify Entailment by Generality . . . 64

5 Evaluating the Performance of our Methodology 67
5.1 Evaluation Scheme . . . 67
5.1.1 Measures to evaluate the performance . . . 67
5.2 All pairs of the Test Set of the first five RTE Challenges . . . 69
5.2.1 All Words . . . 69
5.2.2 Without Stop Words . . . 72
5.2.3 With Multiword Units . . . 75
5.2.4 Summary . . . 78
5.3 Corpus TE by Generality . . . 81
5.3.1 All Words . . . 82

5.3.2 Without Stop Words 84
5.3.3 With Multiword Units 86
5.3.4 Summary 87
5.4 Corpus TE by Generality translated into Portuguese 89
5.4.1 All Words 90
5.4.2 Without Stop Words 91
5.4.3 With Multiword Units 93
5.4.4 Summary 95
5.5 Qualitative Analysis 97

6 Conclusion and Future Work **99**
6.1 Recapitulation 99
6.2 Future Research 101

References **103**

Appendices **113**

Stop Words Lists **115**
A.1 Stop Words List in English 115
A.2 Stop Words List in Portuguese 117

Multiword Units - Extraction of 2-ary Textual Associations **119**
B.1 Multiword Units in English 119
B.2 Multiword Units in Portuguese 136

Sample form submitted to the *"Turkers"* in CrowdFlower **161**

Sample Web Frequencies for Calculations **163**
D.1 All Words 163
D.2 Without Stop Words 166
D.3 With MultiWords Units 168

List of Figures

1.1 Venn diagram of the entailment relation. 8
1.2 Example the entailment by generality relation. 10

3.1 Mechanical Turk process. 46
3.2 Job Calibration Settings in CrowdFlower. 48
3.3 Add Funds in CrowdFlower. 48
3.4 All Reports in CrowdFlower. 49

4.1 Sample calc with all words. 65
4.2 Sample calc with list of the stop words 65
4.3 Sample calc with Multiword Units. 66

List of Tables

2.1 Examples of $T - H$ pairs in RTE-1 (Dagan *et al.*, 2005) 21
2.2 Examples o $T - H$ pair in RTE-2 (Ido *et al.*, 2006) 24
2.3 Examples o $T - H$ pair in RTE-3 (Giampiccolo *et al.*, 2007) 26
2.4 Examples o $T - H$ pair in RTE-4 (Giampiccolo *et al.*, 2008) 28
2.5 Examples o $T - H$ pair in RTE-5 (Bentivogli *et al.*, 2009) 30
2.6 RTE - 1 to RTE - 5 data sets 32
2.7 Best Results in RTE-1 (Dagan *et al.*, 2005) 34
2.8 Best Results in RTE-2 (Ido *et al.*, 2006) 35
2.9 Best Results in RTE-3 (Giampiccolo *et al.*, 2007) 37
2.10 Best Results in RTE-4 (Giampiccolo *et al.*, 2008) 39
2.11 Best Results in RTE-5 41
2.12 Average the top five results 42

3.1 Summary of RTE by Generality corpus annotation task 50

4.1 Web frequencies for calculations with ***All Words*** 59
4.2 Web frequencies for calculations with ***All Words*** 59

5.1 Contingency table for evaluating a binary classifier. For example, a is the number of objects in the category of interest that were correctly assigned to the category. (Manning & Schütze, 1999) 68
5.2 Accuracy Average by RTE Challenges | With All Words 70
5.3 PRECISION - ENTAILMENT Average by RTE Challenges | With All Words 71
5.4 PRECISION - NO ENTAILMENT Average by RTE Challenges | With All Words 72
5.5 Accuracy Average by RTE Challenges | Without Stop Words 73
5.6 PRECISION - ENTAILMENT Average by RTE Challenges | Without Stop Words 74
5.7 PRECISION - NO ENTAILMENT Average by RTE Challenges | Without Stop Words 75
5.8 Accuracy Average by RTE Challenges | With MWU 76
5.9 PRECISION - ENTAILMENT Average by RTE Challenges | With MWU 77
5.10 PRECISION - NO ENTAILMENT Average by RTE Challenges | With MWU 78

5.11 Accuracy Averages | Measures versus Approach 79
5.12 PRECISION - ENTAILMENT Averages | Measures versus Approach 80
5.13 PRECISION - NO ENTAILMENT Averages | Measures versus Approach 81
5.14 Confusion Matrix for all AAM | All Words 82
5.15 Accuracy and Precision by AAM | All Words 83
5.16 Confusion Matrix for all AAM | Without Stop Words 84
5.17 Accuracy and Precision by AAM | Without Stop Words 85
5.18 Confusion Matrix for all AAM | With MWU 86
5.19 Accuracy and Precision by AAM | All Multiword Units (MWU) 87
5.20 Accuracy by AAM 88
5.21 Precisions by AAM 89
5.22 Confusion Matrix for all AAM | All Words 90
5.23 Accuracy and Precision by AAM | All Words 91
5.24 Confusion Matrix for all AAM | Without Stop Words 92
5.25 Accuracy and Precision by AAM | Without Stop Words 93
5.26 Confusion Matrix for all AAM | With MWU 94
5.27 Accuracy and Precision by AAM | All MWU 95
5.28 Accuracy by AAM 96
5.29 Precisions by AAM 97

A.1 Stop Words List in English 115
A.2 Stop Words List in Portuguese 117

B.1 MWU extracted from the first five RTE dataset test. 119
B.2 MWU extracted from the first five RTE dataset test, translated into Portuguese. . . . 136

D.1 Web frequencies for calculations with ***All Words*** 163
D.2 Web frequencies for calculations with ***All Words*** 164
D.3 Web frequencies for calculations without ***Stop Words*** 166
D.4 Web frequencies for calculations without ***Stop Words*** 166
D.5 Web frequencies for calculations with ***MultiWords Units*** 168
D.6 Web frequencies for calculations with ***MultiWords Units*** 168

List of Abbreviations

AAM	Asymmetric Association Measures.
AC	Accuracy.
AI	Artificial Intelligence.
AIS	Asymmetric InfoSimba Similarity.
AISs	Simplified Asymmetric InfoSimba Similarity.
ATN	Augmented Transition Network.
CD	Comparable Documents.
CM	Confusion Matrix.
DIRT	Discovery of Inference Rules from Text.
DUC	Document Understanding Conferences.
HITs	Human Intelligence Tasks.
IAM	Informative Asymmetric Measure.
IE	Information Extraction.
IR	Information Retrieval.
IS	InfoSimba Similarity.
ME	Mutual Expectation.
ML	Machine Learning.
MT	Machine Translation.
MTurk	Amazon Mechanical Turk.
MWU	Multiword Units.
NE	Named Entity.
NIST	National Institute of Standards and Technology.
NL	Natural Language.
NLP	Natural Language Processing.
P	Precision.
PASCAL	Pattern Analysis, Statistical Modelling and Computational Learning.
PO	Precision-Oriented.
PP	Paraphrase Acquisition.

QA	Question Answering.
RAIS	Recursive Asymmetric InfoSimba Similarity.
RC	Reading Comprehension.
RTE	Recognizing Textual Entaiment.
SUM	Summarization.
TAC	Text Analysis Conference.
TE	Textual Entaiment.
WWW	World Wide Web.

CHAPTER 1

INTRODUCTION

"Victory is always possible for the person
who refuses to stop fighting."
Napoleon Hill

In this Chapter we introduce the context and the motivations underlying the present research work, we analyze the notion and variants of entailment and consequently Textual Entaiment (TE). Also here we introduce our objective in this work - identifying entailment by generality in pairs of sentences.

1.1 Natural Language Processing

We can not imagine a world without communication. Every living being must communicate to survive. For us, human beings, language is a fundamental aspect and it is a crucial component of our life. In written form it serves as a long-term record of knowledge from one generation to the next. In spoken form it serves as our primary means of coordinating our day-to-day behavior with others. Thus, producing language is above all a social activity.

NLP is a field of computer science and linguistics concerned with the interactions between computers and humans by means of natural language. In theory, NLP is a very attractive method of human-computer interaction. Natural Language (NL) understanding is sometimes referred to as an Artificial Intelligence (AI) complete problem, because NL recognition seems to require extensive knowledge about the outside world and the ability to manipulate it. NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence.

Modern NLP algorithms are grounded in Machine Learning (ML), especially statistical ML. Research into modern statistical NLP algorithms requires understanding of a number of fields, including linguistics, computer science, statistics (particularly Bayesian Statistics), linear algebra and optimization theory.

1.1.1 Historical Review

Work in the NLP field has concentrated first on one problem, then on another, sometimes because solving problem X depends on solving problem Y but sometimes just because problem Y seems more tractable than problem X, or because there is market interest in a solution to Y. There has been very substantial progress, both in understanding how to do NLP and in actually doing it, since work in the field took off in the 1950s. In the last twenty-five years in particular, advances in computing technology have made it possible to implement ideas that could only be adumbrated before, to consolidate research, and to carry speech and language processing into the ordinary world. Sometimes the scientific advance in NLP, or the computational linguistics underlying it, is less than the onward rush of information technology field evident in the fifty-year period reviewed here.

Sometimes innovation is only old ideas reappearing in new guises, like lexical list approaches to NLP, or shallow parsing. But the new costumes are better made, of better materials, as well as more becoming: so the research is not so much going round in circles as ascending a spiral.

The work of the late 1940s to late 1960s, was focused on Machine Translation (MT). Following a few early birds, including Booth and Richens' investigations and Weaver's influential memorandum on translation of 1949, research on NLP began in earnest in the 1950s. Automatic translation from Russian to English, in a very rudimentary form and limited experiment, was exhibited in the IBM-Georgetown Demonstration of 1954 (Hutchins *et al.*, 1955). The journal *Mechanical Translation*, the ancestor of *Computational Linguistics*, also began publication in 1954. The first international conference on Mechanical Translation was held in 1952, the second in 1956 (the year of the first AI conference); at the important Washington International Conference on Scientific Information of 1958 language processing was linked with information retrieval, for example in the use of a thesaurus, Minsky drew attention to AI, and Luhn provided auto-abstracts (actually extracts) for one session's papers. The Teddington International Conference on Machine Translation of Language and Applied Language Analysis in 1961 was perphars the high point of this phase: it reported work done in many countries on many aspects of NLP including morphology, syntax and semantics, in interpretation and generation, and ranging from formal theory to hardware.

Some notably successful NLP systems developed in the 1960s were SHRDLU[1] - *"SHRDLU was primarily a language parser that allowed user interaction using English terms. The user instructed SHRDLU to move various objects around in a "blocks world" containing various basic objects: blocks, cones, balls, etc. What made SHRDLU unique was the combination of four simple ideas that added up to make the simulation of "understanding" far more convincing."*[2] - a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA (Weizenbaum, 1966), a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum[3] between 1964 to 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly

[1] `http://hci.stanford.edu/~winograd/shrdlu/` [Last access: 14th December, 2013]

[2] `http://en.wikipedia.org/wiki/SHRDLU` [Last access: 14th December, 2013]

[3] `http://en.wikipedia.org/wiki/Joseph_Weizenbaum` [Last access: 14th December, 2013]

human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?".

In 1969 Roger Schank[1] introduced the conceptual dependency theory for natural language understanding (Schank & Tesler, 1969).

In 1970, William A. Woods introduced the Augmented Transition Network (ATN) to represent natural language input (Woods, 1970). Instead of phrase structure rules ATN used an equivalent set of finite state automata that were called recursively. ATN and their more general format called "generalized ATN" continued to be used for a number of years. During the 70's many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data.

Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. This was due both to the steady increase in computational power resulting from Moore's Law[2] and the gradual lessening of the dominance of Chomskyan[3] theories of linguistics (e.g. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing. Some of the early machine learning algorithms, such as decision trees, produced systems of hard *if-then* rules similar to existing hand-written rules. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to the features making up the input data. The cache language models upon which many speech recognition systems now rely are examples of such statistical models. Such models are generally more robust when given unfamiliar input, especially input that contains errors (as is very common for real-world data), and produce more reliable results when integrated into a larger system comprising multiple subtasks.

Many of the notable early successes occurred in the field of MT, due especially to work at IBM Research, where successively more complicated statistical models were developed. These systems were able to take advantage of existing multilingual textual corpora that had been produced by the Parliament of Canada[4] and the European Union[5] as a result of laws calling for the translation of all governmental proceedings into all official languages of the corresponding governmental systems. However, most other systems depended on corpora developed for specific task, which was (and often continues to be) a major obstacle to the success of these systems. As a result, a great deal of research has gone into methods of more effectively learning from limited amounts of data.

Recent research has increasingly focused on unsupervised (our case) and semi-supervised learning

[1] `http://en.wikipedia.org/wiki/Roger_Schank` [Last access: 14th December, 2013]

[2] `http://en.wikipedia.org/wiki/Moore's_Law` [Last access: 14th December, 2013]

[3] `http://web.mit.edu/linguistics/people/faculty/chomsky/index.html` [Last access: 14th December, 2013]

[4] `http://www.parl.gc.ca/common/index.asp?Language=E` [Last access: 14th December, 2013]

[5] `http://europa.eu/index_en.htm` [Last access: 14th December, 2013]

algorithms. Such algorithms are able to learn from data that has not been hand-annotated with the desired answers, or using a combination of annotated and raw data. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the World Wide Web (WWW)), which can often make up for the inferior results.

1.1.2 Applications of Natural Language Processing

NLP is an interdisciplinary research area at the border between linguistics and AI aiming at developing computer programs capable of human-like activities related to understanding or producing texts or speech in a natural language, such as English or Chinese.
The most important applications of NLP include Information Retrieval (IR) and information organization, MT, and natural language interfaces, Information Extraction (IE), Summarization (SUM), search engine, among others. However, as in any science, the activities of the researchers are mostly concentrated on its internal art and craft, that is, on the solution of the problems arising in analysis or generation of natural language text or speech, such as syntactic and semantic analysis, disambiguation, or compilation of dictionaries and grammars necessary for such analysis.

1.2 Motivation and Rationale

Natural Language (NL) allows the same meaning to be expressed in many different ways, making automatic understanding particularly challenging. Almost all computational linguistics tasks such as IR, Question Answering (QA), IE, SUM and MT have to cope with this phenomenon.
Inference is generally perceived as the process by which new knowledge is inferred from given information. For example, the Merriam-Webster online dictionary[1] defines the first sense of infer as *"to derive as a conclusion from facts or premises"*. Somewhat more technically, inference is defined as *"the act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former"*.
Moving to the realm of NLP, we can analogically perceive inference over information stated in human language. Such inference can be defined as the process of concluding the truth of a textual statement based on (the truth of) another given piece of text. This language-oriented view on inference was captured by the *textual entailment* paradigm, originally proposed by Dagan & Glickman (2004) and subsequently established through the series of benchmarks known as the *PASCAL Recognising Textual Entailment (RTE) Challenges*.
While capturing a generic notion of inference over texts, the introduction of entailment recognition as a computational task was particularly motivated by its overarching potential for NLP applications.

[1] `http://www.merriam-webster.com/` [Last access: 14th December, 2013]

For example, consider a QA scenario, addressing the question *"Who painted 'The Scream'?"*. In order to provide the answer *"Edvard Munch"*, based on the text snippet *"Norway's most famous painting, 'The Scream' by Edvard Munch,…"*, the QA system needs to validate that the hypothesized answer statement *"Edvard Munch painted 'The Scream'."* is indeed entailed (inferred) by the given text.
Entailment is widely used in many aspects of the human life. Assume that someone is seeking for something and he or she searches for the answer from books, friends, or the Web. In most cases, the information gathered or retrieved is not the exact answer, although the (information) seeker may have one in his or her mind. Instead, the consequences of the original goal may be detected, so the entailment plays a role and confirms or denies the original information being sought (Dagan *et al.*, 2013).
For example, John wants to know whether the Amazon river is the longest river in the world. Naturally, he can find the exact lengths of the Amazon and other rivers he knows of, and then compare them. But once he sees *"Egypt is one of the countries along the longest river on earth"*, he can already infer that Amazon is not the longest river, since Egypt and the Amazon river are not on the same continent. Similarly, assuming that Albert is not sure who is the current president of the U.S., Bush or Obama, since both *"president Bush"* and *"president Obama"* are retrieved. If he performs an inference based on one of the retrieved documents containing *"George Bush in retirement"*, the answer is obvious. In short, finding out the exact information is not always trivial, but inference can help a lot. In both cases, the retrieved information entails the answer instead of being the precise answer.
Entailment also occurs frequently in our daily communication, with respect to language understanding and generation. Usually we do not literally interpret each other's utterances, nor express ourselves in a straight way. For example:

- Tom: *Have you seen my iPad?*
- Robin: *Oh, nice! I'd like to have one too.*
- Tom: *You have to get one.*

The dialogue seems to be incoherent, if we literally and individually interpret each sentence. Firstly, Tom asks a yes-no question, but Robin does not directly give the answer. Instead, Robin implies that he has not seen it before the conversation by showing his compliment to it (*"Oh, nice!"*). Probably Tom is showing his iPad to Robin during the conversation. Robin's second sentence also implies that he does not have an iPad till then, and therefore Tom's response is a suggestion for him to get one.
If we literally interpret the conversation, it sounds a bit awkward. Here is one possibility:

- Tom: *Here is my iPad.*
- Robin: *I haven't seen it before. It is nice. I don't have one, but I'd like to have one.*
- Tom: *I suggest you get one.*

Although the interpreted version may be easier for the computers to process human dialogues, the original conversation occurs more naturally in our daily life. Each utterance in the interpreted version is actually implied or entailed by the utterances in the original conversation. Consequently, if we want to build a dialogue system, dealing with this kind of implication or entailment is one of the key challenges. Let alone there is common sense knowledge which does not appear in the dialogue but is nevertheless acknowledged by both speakers, e.g., what an iPad is.
RTE was proposed by Dagan & Glickman (2004) as a generic NLP task in order to overcome the problem of lexical, syntactic and semantic variability in natural languages. In 2005, the RTE Challenge has been launched by Dagan *et al.* (2005), defining TE as a task for automatic systems.
Given a text T and a hypothesis H, the task consists of deciding whether the meaning of H can be inferred from the meaning of T. The following examples show $T - H$ pairs for which the entailment relation holds (**Example 1**) or not (**Example 2**):

- **Example 1**
  T: *Euro-Scandinavian media cheer Denmark vs Sweden draw.*
  H: *Denmark and Sweden tie.*
  Entailment: YES
- **Example 2**
  T: *Oracle had fought to keep the forms from being released.*
  H: *Oracle released a confidential document.*
  Entailment: NO

In the many evaluation campaigns that in recent years addressed the TE recognition problem, complex definitions of the task have been proposed. The released datasets reflect the long-term objective of creating more natural evaluation settings. These include the formulation of TE as a search task[1] (i.e. finding all the sentences in a set of documents that entail a given hypothesis), the use of TE to approach the Answer Validation Exercise[2] (emulate human assessment of QA responses and decide whether an answer to a question is correct or not according to a given text), and the very recent effort to explore multi-directional TE recognition[3] (moving from YES/NO to directional entailment judgements such as Forward, Backward and Bidirectional). Consequently, a large number

[1] RTE: `http://www.nist.gov/tac/2010/RTE/` [Last access: 14th December, 2013]
[2] AVE: `http://nlp.uned.es/clef-qa/` [Last access: 14th December, 2013]
[3] NTCIR-9 RITE: `http://artigas.lti.cs.cmu.edu/rite/` [Last access: 14th December, 2013]

of methods and resources for TE has been published or released.

As for the NLP perspective, RTE can be viewed as a generic semantic processing module, which serves for other tasks. For instance, it has already been successfully used for question answering (Harabagiu & Hickl, 2006), including answer validation (Peñas *et al.*, 2008; Rodrigo *et al.*, 2009), information extraction (Roth *et al.*, 2009), and MT evaluation (Padó *et al.*, 2009a). In the long term, RTE can also play an important role in understanding conversation dialogues (Zhang & Chai, 2010), metaphors (Agerri, 2008), and even human-robot communication (Bos & Oka, 2007).

Given the multiple applicability that Textual Entailment can have, we understand that there are several types of implications, where each type of implication stems or suits specific task. Proof of this is the diversity of methodologies and results presented in the RTE challenges. It is accepted that textual entailment is not an exact science and we believe that there is still much to be investigated in this area.

In this thesis we introduce a new concept, ***Entailment by Generality***. This new paradigm can be defined as the relation that holds between a specific statement that implies a more general one, for example, *strawberry* (specific) implies *fruit* (general), because *strawberry* is really a *fruit*, but *fruit* does not necessarily imply *strawberry*, because, *fruit* can be *strawberry* but can also be *orange*, *banana*, or other *fruit*.

Also in this thesis we present our methodology - unsupervised, language-independent and threshold free - for learning to identify entailment by generality between two sentences. This technology is enabling **Ephemeral Clusters Summarization of Web Pages** (Dias *et al.*, 2011), useful for optimized **Search Engine** results visualization.

In the context of Ephemeral Clustering of Web Pages, it can be interesting to label each cluster with a small summary instead of just a label. Thus we are interested to find the best web snippet, which summarizes and subsumes all the other web snippets within an ephemeral cluster. This summary can be defined as a general entailment from a specific information characteristic of the cluster.

Although, Ephemeral Clustering has been studied for more than a decade, it has received low user acceptance. According to us, there are two main reasons for this situation. First, state-of-the-art systems tend to generate an excessive number of clusters. As a consequence, browsing through a high number of clusters is mostly similar to searching through a high number of Web pages. Second, improved user interfaces can only be achieved through high quality cluster labeling. In the optimal case, the labels of the clusters should clearly evidence their overall contents. However, very little has been proposed in the community to overcome the latter obstacle. The only exception is certainly (de Buenaga *et al.*, 2008) who propose to increase the expressiveness of cluster labels with a summary obtained by classical Multi-document Summarization techniques. However, their solution is fulltext based and can not be applied in real-time real-world applications. As a consequence, we propose to increase cluster expressiveness based on finding the web snippet within the ephemeral cluster, which best summarizes and subsumes all the other web snippets present in the cluster. For that purpose, we propose a different methodology based on TE by Generality.

1.3 Linguistic Notion of Entailment

The application-oriented notion of textual entailment is related, of course, to a classical logic-based notion of entailment in linguistics. A common definition of entailment in formal semantics 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. However, the TE definition allows for cases in which the truth of the hypothesis is highly plausible ("most likely true"), for most practical purposes, rather than certain.

In propositional and predicate logic, entailment (or logical implication) describes a relation between one sentence or a set of sentences - the entailing expressions - represented as formulae of a formal language, and another sentence that is entailed. Formally, given a set of formulae $\Gamma = A_1, ..., A_n$ and a formula B, we say that Γ semantically entails $B(\Gamma \models B)$ if and only if every model (or interpretation) of $A_1, ..., A_n$ is also a model of B. The Venn diagram of this relationship is show in Figure 1.1.

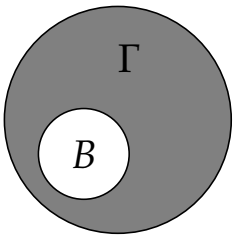

Figure 1.1: Venn diagram of the entailment relation.

1.3.1 Variants of the Entailment

As already mentioned in this thesis, we argue that there are several types of entailment, for example in the study done in (Pazienza *et al.*, 2005),they present three types of entailment can be defined:

1. *Semantic Subsumption* - T and H express the same fact, but the situation described in T is more specific than the situation in H. The specificity of T is expressed through one or more semantic operations. For example in the sentential pair:
   - H: The cat eats the mouse.
   - T: The cat devours the mouse.

   T is more specific than H, as eat is a semantic generalization of devour.

2. *Syntactic Subsumption* - T and H express the same fact, but the situation described in T is more specific than the situation in H. The specificity of T is

expressed through one or more syntactic operations. For example in the pair:

- H: The cat eats the mouse.
- T: The cat eats the mouse in the garden.

T contains a specializing prepositional phrase.

3. *Direct Implication* - H expresses a fact that is implied by a fact in T. For example:

- H:The cat killed the mouse.
- T: The cat devours the mouse.

H is implied by T, as it is supposed that killed is a precondition for devour. In (Dagan & Glickman, 2004) syntactic subsumption roughly corresponds to the restrictive extension rule, while direct implication and semantic subsumption to the axiom rule.

In (Pazienza *et al.*, 2005) despite the two types of subsumption entailment, direct implication underlies deeper semantic and discourse analysis. In most cases, as implication concerns two distinct facts in T and H, and as facts are usually expressed through verbs, it follows that the implication phenomenon is strictly tied to the relationship among the T and H verbs. In particular, it is interesting to notice the temporal relation between T and H verbs, as described in (Miller, 1995). The two verbs are said to be in temporal inclusion when the action of one verb is temporally included in the action of the other (e.g. snore $->$ sleep). Backward-presupposition stands when the H verb happens before the T verb (win entails play). In causation a stative verb in H necessarily follows a verb of change in T (e.g. give $->$ have). In this case, the temporal relation is thus inverted with respect to backward-presupposition. Such considerations leave space to the application of temporal and verb analysis techniques both in the acquisition and recognition tasks.

Ultimately, we want to regard entailment by generality as a relation between utterances (that is, sentences in context), where the context is relevant to understand the meaning. Then, considering study in (Pazienza *et al.*, 2005), we understand that the relation entailment by generality can be compared with one of three relations:

- *Semantic Subsumption*;
- *Syntactic Subsumption*;
- Or a combination - *Semantic Subsumption* $+$ *Syntactic Subsumption*;

For us, in the most common definition, Entailment by Generality can be defined as the entailment from specific sentence towards a more general sentence. (Dias *et al.*, 2011; Pais *et al.*, 2011).

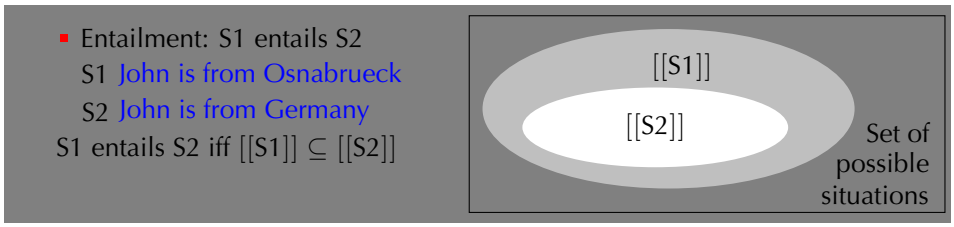

Figure 1.2: Example the entailment by generality relation.

1.4 Context of Textual Entailment

Natural languages allow to express the same meaning in many possible ways, making automatic understanding particulary challenging. Almost all computational linguistics tasks such as IR, QA, IE, text summarization and MT have to cope with this phenomenon.

Within TE framework, a text T is said to entail a textual hypothesis H if the truth of H can be inferred from T. This means that most people would agree that the meaning of T implies that of H. Somewhat more formally, we say that T entails H when some representation of H can be "matched" with some (or part of a) representation of T, at some level of granularity and abstraction.

Dagan & Glickman (2004) define TE as a relationship between a coherent textual fragment T and a language expression, which is considered as a hypothesis H. Entailment holds (i. e. $T \rightarrow H$) if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. This relationship is directional and asymmetric, since the meaning of one expression may usually entail the other, while entailment in the other direction is less certain.

This definition of textual entailment captures quite broadly the reasoning about language variability needed by different applications aimed at natural language understanding and processing (Androutsopoulos & Malakasiotis, 2010; Dagan *et al.*, 2009). For instance, a QA system has to identify texts that entail the expected answer. Given the question "Who painted the Mona Lisa?", the text "Among the works created by Leonardo da Vinci in the 16th century is the small portrait known as the Mona Lisa or la 'Gioconda'", entails the expected answer "Leonardo da Vinci painted the Mona Lisa". Similarly, in IR relevant documents should entail the combination of semantic concepts and relations denoted by the query. In IE, entailment holds between different text variants expressing the same target relation (Romano *et al.*, 2006). In text summarization, an important processing stage is sentence extraction, which identifies the most important sentences of the texts to be summarized; especially when generating a single summary from several documents (Barzilay & McKeown, 2005), it is important to avoid selecting sentences that convey the same information as other sentences that have already been selected (i.e. that entail such sentences). Also in MT, an entailment relation should hold:

1. among machine-generated translations and human-authored ones that may use different phrasings in the evaluation phase (Padó *et al.*, 2009b), or

2. in the translation phase, between source language words and longer phrases that have not been encountered in training corpora (Mirkin *et al.*, 2009).

Other applications that could benefit from such inference model are reading comprehension systems (Nielsen *et al.*, 2009).

Below, we give a few variants of informal definitions for textual entailment.

- Dagan *et al.* (2005) - *[...] a text T entails a hypothesis H if, typically, a human reading T would infer that H is most likely true*;

- A definition of entailment in formal semantics (Chierchia & McConnell-Ginet, 2000) reads - *A text T entails another text H if H is true in every circumstance (possible world) in which T is true.*

Several definitions are given by the participants in various RTE challenges:

- *T entails H if we have a sequence of transformations applied to T such that we can obtain H with an overall cost below a certain threshold, empirically estimated on the training data (Kouylekov & Magnini, 2005)*;

- If the BLEU's output is higher than a threshold value the entailment is marked as TRUE, otherwise as FALSE (Pérez & Alfonseca, 2006);

- T entails H if we succeed to extract a maximal subgraph of XDG_T that is in isomorphism relation with a subgraph XDG_H (Pazienza & Pennacchiotti, 2005);

- In Guidelines of RTE-4[1] Challenge - *T entails H if the truth of H can be inferred from T within the context induced by T.*

1.4.1 Probabilistic Textual Entailment

In many intuitive cases, the textual entailment recognition task may be perceived as being deterministic (Glickman & Dagan, 2005). For example, given the hypothesis h_1 = *"Harry was born in Iowa"* and a candidate text t_1 that includes the sentence *"Harry's birthplace is Iowa"*, it is clear that t_1 does (deterministically) entail h_1, and humans are likely to have high agreement regarding

[1] `http://www.nist.gov/tac/2008/rte/rte.08.guidelines.html` [Last access: 14th December, 2013]

this decision. In many other texts, though, entailment inference is uncertain and has a probabilistic nature. For example, a text t_2 that includes the sentence *"Harry is returning to his Iowa hometown to get married."* does not deterministically entail the above h_1 since Harry might have moved to Iowa as a child. Yet, it is clear that t_2 does add substantial information about the correctness of h_1. In other words, the probability that h_1 is indeed true given the text t_2 ought to be significantly higher than the prior probability of h_1 being true. More specifically, we might say that the probability p of h_1 being true should be estimated based on the percentage of cases in which someone's reported hometown is indeed his/her birthplace. Accordingly, we wouldn't accept t_2 as a definite assessment for the truth of h_1. However, in the absence of other definite information, t_2 may partly satisfy our information need for an assessment of the probable truth of h_1, with p providing a confidence probability for this inference.

Meanings are captured in Glickman & Dagan (2005) model by hypotheses and their truth values. Let T denotes a space of possible texts, and $t \in T$ a specific text and 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. For now it is assumed that h is represented as a textual statement, but in principle other representations for h may fit their framework as well. A semantic state of affairs is captured by a possible world w: $H \rightarrow 0,1$, which is defined as a mapping from H to $0 = False, 1 = True$, representing the set of w's concrete truth value assignments for all possible propositions. Accordingly, W denotes the set of all possible worlds.

Glickman & Dagan (2005) present a first attempt to define a generative probabilistic setting for TE, which allows a clear formulation of probability spaces and concrete probabilistic models for this task. According to their definition, a text t probabilistically entails a hypothesis h $(t \rightarrow h)$ if t increases the likelihood of h being true, i.e. if $P(Tr_h = 1|t) > P(Tr_h = 1)$, where Tr_h is the random variable whose value is the truth value assigned to h in a given world.

From this applied empirical perspective, textual entailment represents therefore an uncertain - but highly plausible - relation, that has a probabilistic nature.

1.4.2 Recognizing Textual Entailment

The RTE task, as defined by Dagan *et al.* (2005), and established in the RTE Challenges, is formulated as follows:

Textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T (the entailing *"Text"*) and H (the entailed *"Hypothesis"*). We say that T entails H if humans reading T would typically infer that H is most likely true.

Basically, RTE is the task of deciding, given two text fragments, whether the meaning of one

of the texts is entailed (can be inferred) from the other text.

As noted by Dagan *et al.* (2005), this definition is based on common human understanding of language, much like the definition of any other language understanding task. Accordingly, it enables the creation of gold-standard evaluation data sets for the task, where humans can judge whether the entailment relation holds for given Text-Hypothesis pairs. This setting is analogous to the creation of gold standards for other text understanding applications like QA and IE, where human annotators are asked to judge whether the target answer or relation can indeed be inferred from a candidate text. The distinguishing characteristic of the textual entailment task is that it captures textual inference in a generic, application-independent manner. This allows research to focus on core inference issues, while making the results applicable across application areas.

Similar to other semantic annotation tasks, such as those mentioned above, the RTE judgment criterion has some fuzziness with respect to *"what a person would typically infer"*, particularly in boundary cases. However, the various RTE annotation efforts have shown that sufficiently consistent human judgments can be obtained, allowing research progress on this task (Dagan *et al.*, 2013).

Also, this task captures generically a broad range of inferences that are relevant for multiple applications. For example, QA system has to identify texts that entail the expected answer. Given the question *"Who is John Lennon's widow?"* the text *"Yoko Ono unveiled a bronze statue of her late husband, John Lennon, to complete the official renaming of England's Liverpool Airport as Liverpool John Lennon Airport"* entails the expected answer *"Yoko Ono is John Lennon's widow"*. Similarly, semantic inference needs of other text-understanding applications such as IR, IE and MT evaluation can be cast as entailment recognition (Candela *et al.*, 2006). A necessary step in transforming textual entailment from a theoretical idea into an active empirical research field was the introduction of benchmarks and an evaluation forum for entailment systems.

1.5 Our Proposal for RTE by Generality

We introduce the paradigm of TE by Generality, which can be defined as the entailment from a specific sentence towards a more general sentence. For example, from sentences (1) and (2) extracted from RTE-1, we would easily state that $(1) \rightarrow (2)$ as their meaning is roughly the same although sentence (2) is more general than sentence (1).

- (1) Mexico City has a very bad pollution problem because the mountains around the city act as walls and block in dust and smog.

- (2) Poor air circulation out of the mountain-walled Mexico City aggravates pollution.

To understand how Textual Entailment by Generality can be modeled for two sentences, we propose a new paradigm based on a new Informative Asymmetric Measure (IAM), called the Asymmetric InfoSimba Similarity (AIS) measure. Instead of relying on the exact matches of words between texts, we propose that one sentence infers the other one in terms of generality if two constraints hold: (a) if and only if both sentences share many related words and (b) if most of the words of a given sentence are more general than the words of the other sentence. As far as we know, we are the first to propose an unsupervised, language-independent, threshold free methodology in the context of TE by Generality, although the approach from Glickman & Dagan (2005) is based on similar assumptions. This new proposal is exhaustively evaluated against the first five RTE datasets by testing different Asymmetric Association Measures (AAM) in combination with the AIS. In particular, the RTE-1 as it is the only dataset for which there exist comparable results with linguistic-free methodologies (Bayer *et al.*, 2005; Glickman & Dagan, 2005; Perez *et al.*, 2005).
Finally, we propose to avoid the definition of a "hard" threshold and study exhaustively asymmetry in language i.e. not just by the conditional probability as done in Glickman & Dagan (2005). For that purpose, we propose a new IAM called the AIS combined with different Association Measures.

1.6 Structure of the Thesis

The Thesis is structured as follows:

- **Chapter 2** gives an overview of the research in TE, in particular, it focuses on the first five RTE Challenges (datasets and annotations, and relevant resources and tools).
- **Chapter 3** describes the methodology for the creation of the corpus of pairs $T->H$, to learn RTE by Generality, taking advantage of crowdsourcing. This corpus was created with the help of the CrowdFlower service[1] which provides a crowdsourcing interface to MTurk[2] for non-US citizens.
- **Chapter 4**, the core of our work - our methodology, this chapter presents an IAM called the Simplified Asymmetric InfoSimba Similarity (AISs), which we combine with different AAM to recognize the specific case of TE by Generality. The AISs provides an unsupervised, language-independent and threshold free solution.
- **Chapter 5** reports extensively several experiments and respective results on three datasets. In this experiments, we extract MWU with SENTA and used the *Stop Words* lists.
- **Chapter 6** concludes the Thesis drawing final remarks and suggesting directions for future improvements.

[1] `http://crowdflower.com/` [Last access: 14th December, 2013]
[2] `http://www.mtruk.com/` [Last access: 14th December, 2013]

- **Appendix A** show *Stop Words* list in English [1] and show other *Stop Words* list in Portuguese [2].
- **Appendix B** show two lists the MWU, in English and Portuguese, prepared for the first five RTE Challenges.
- **Appendix C** illustrates the submitted form to *"Turkers"* in *CrowdFlower*.
- **Appendix D** presents the Web frequencies for the calculations in the pair $T - H$ in the section 4.5.

[1] Source: `http://www.microsoft.com/en-us/download/confirmation.aspx?id=10024` [Last access: 14th December, 2013]

[2] Source: `http://www.linguateca.pt/chave/stopwords/` [Last access: 14th December, 2013]

CHAPTER 2

RELATED WORK

"Never regard study as a duty,
but as the enviable opportunity to learn to know the liberating influence
of beauty in the realm of the spirit for your own personal joy and
to the profit of the community to which your later work belongs."
Albert Einstein

This chapter presents the state of the art of the research in TE. Given the significant number of publications on this topic, we focus on a set of relevant works that are unsupervised and language-independent.

2.1 Overview of the First Five RTE Challenges

The first three RTE competitions - RTE-1[1], RTE-2[2] and RTE-3[3] - were organized by Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) Network[4].
The 2008 and 2009 (RTE-4[5] and RTE-5[6], respectively) challenges were organized within the Text Analysis Conference (TAC). The TAC is a new series of evaluation workshops organized to encourage research in NLP and related applications, by providing a large test collection, common evaluation procedures, and a forum for organizations to share their results. Year-to-year, new features were added in every new competition.
In 2005, the RTE Challenge was launched by Dagan *et al.* (2005), defining TE as a task for automatic systems. Given two texts T and H, the task consists in deciding whether the meaning of H can be inferred from the meaning of T. The following example shows a $T - H$ pair for which the entailment

[1] `http://pascallin.ecs.soton.ac.uk/Challenges/RTE/` [Last access: 14th December, 2013]
[2] `http://pascallin.ecs.soton.ac.uk/Challenges/RTE2` [Last access: 14th December, 2013]
[3] `http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/` [Last access: 14th December, 2013]
[4] `http://www.pascal-network.org` [Last access: 14th December, 2013]
[5] `http://www.nist.gov/tac/2008/rte/index.html` [Last access: 14th December, 2013]
[6] `http://www.nist.gov/tac/2009/RTE/index.html` [Last access: 14th December, 2013]

relation holds:

- T: In the end, defeated, Antony committed suicide and so did Cleopatra, according to legend, by putting an asp to her breast.
- H: Cleopatra committed suicide.

At present, TE is considered an interesting and challenging topic within the NLP community, due to its many potential applications. The PASCAL Network promoted a generic evaluation framework covering semantic-oriented inferences for several NLP applications, which led to launch the RTE Challenge. Many research areas such as IE, QA, IE, text summarization and MT have to cope with different kinds of inference mechanisms, closely related to the entailment notion. In this direction, some works attempted to apply textual entailment to various NLP tasks in order to benefit from a semantic inference framework, and to potentially improve their performances (Glickman, 2009).

2.1.1 Datasets and Annotations

2.1.1.1 RTE-1

The set of $Text - Hypothesis$ pairs used in the first RTE challenge was collected by human annotators. It consists of seven subsets, which correspond to typical success and failure settings in other applications. Within each application setting the annotators selected an equal number of both positive entailment examples, where T is judged to entail H, and negative examples, where entailment does not hold. Typically, T consists of one sentence (sometimes two) while H was most often made of a single short sentence. Part of the examples were collected using external sources, such as available datasets or systems as follows:

- Document Understanding Conferences (DUC) 2004 MT evaluation data, from the National Institute of Standards and Technology (NIST)[1];
- TextMap Question Answering online demo, from the Information Sciences Institute[2];
- Relation Recognition dataset, from University of Illinois at Urbana-Champaign[3];
- DIRT paraphrase database (online demo), from the University of Southern California[4];
- The output of the TEASE system for extracting entailment relations and paraphrases (Szpektor *et al.*, 2004);

[1] `http://duc.nist.gov/duc2004/` [Last access: 14th December, 2013]
[2] `http://www.textmap.com/` [Last access: 14th December, 2013]
[3] `http://l2r.cs.uiuc.edu/~cogcomp/` [Last access: 14th December, 2013]
[4] `http://demo.patrickpantel.com/demos/lexsem/paraphrase.htm` [Last access: 14th December, 2013]

- Corpus of aligned sentences extracted from monolingual comparable corpora, Columbia University[1].

A fraction of the examples were collected from the Web, focusing on the general news domain. In all cases the decision as to which example pairs to be included 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. It is interesting to note that more negative examples than positive ones were produced in the cases where T and H have a very high degree of lexical overlap (Dagan *et al.*, 2005). Bellow are listed the specific routines followed for each application area.

- **Collecting Information Retrieval (IR) pairs**: 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. The queries were selected by examining prominent sentences in news stories, and then submitted to a web search engine. Candidate texts T were selected from the search engine's retrieved documents, picking candidate texts that either do or do not entail the hypothesis;

- **Collecting Comparable Documents (CD) pairs**: Annotators identified $T - H$ pairs by examining a cluster of comparable news articles that cover a common story. They examined pairs of aligned sentences that overlap lexically, in which semantic entailment may or may not hold. Some pairs were identified on the web using *Google News*[2] and others were taken from a corpus of aligned sentences. 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;

- **Collecting Reading Comprehension (RC) pairs**: This task corresponds to a typical reading comprehension exercise in human language teaching, where students are asked to judge whether a particular statement can be inferred from a given text story. Annotators were asked to create such $T - H$ pairs, that constitute an adequate reading comprehension test for high school students;

- **Collecting Question Answering (QA) pairs**: The TextMap Web Based Question Answering system, available online, was queried with questions taken from CLEF-QA and TREC, also the annotators were allowed to construct their own questions. For a given question, the annotators chose first a relevant text snippet T that was suggested by the system as including the

[1] `http://www.cs.columbia.edu/~noemie/alignment/` [Last access: 14th December, 2013]
[2] `http://news.google.com/` [Last access: 14th December, 2013]

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 the question, "Who is Ariel Sharon?" and taking a candidate answer text "Israel's Prime Minister, Ariel Sharon, visited Prague." as T, the hypothesis H is formed by turning the question into the statement "Ariel Sharon is Israel's Prime Minister.", thus producing a *True* entailment pair;

- **Collecting Information Extraction (IE) pairs**: For this task the annotators used an available dataset annotated for the IE relations "kill" and "birth place" produced by University of Illinois at Urbana-Champaign, as well as general news stories in which they identified manually "typical" IE relations. Given an IE relation of interest, annotators choose T candidate among news story sentences in which the relation holds. As a hypothesis they created a straightforward formulation of the IE relation. 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, thus producing a *True* entailment pair;

- **Collecting Machine Translation (MT) pairs**: Two translations of the same text, an automatic translation and a gold standard human translation, were compared and modified in order to obtain $T - H$ pairs. The automatic translation was alternately taken as either T or H, where a correct translation corresponds to *True* entailment. The automatic translations were grammatically adjusted whenever needed;

- **Collecting Paraphrase Acquisition (PP) pairs**: PP systems attempt to generate pairs of expressions that convey mostly equivalent or entailing meanings being at the same time as much grammatically correct as possible. Annotators selected a text T from some news story which includes a certain relation, for which a paraphrase acquisition system produced a set of paraphrases. Correct paraphrases suggested by the system yielded *True* $T - H$ pairs; otherwise a *False* example was generated.

In a second phase of this dataset production process the examples produced by one annotator were validated by the other annotator who received only the text and hypothesis pair, without any additional information from the original context. The annotators agreed in their judgment for roughly 80% of the examples, which corresponded to a 0.6 Kappa level (moderate agreement). The 20% of the pairs on which the judges disagreed were discarded. A third person reviewed the remaining examples and eliminated about additional 13% of the original examples, which seemed controversial.

The final dataset is believed to represent a broad range of naturally occurring entailment factors. However, it is unclear whether it corresponds to a particular representative distribution of these factors. 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. A sample of this dataset is given in Table 2.1.

ID	TEXT	Hypothesis	TASK	ENTAILMENT
1	iTunes software has seen strong sales in Europe.	Strong sales for iTunes in Europe.	IR	True
2	Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.	The Beatles perform at Cavern Club at lunchtime.	IR	True
3	American Airlines began laying off hundreds of flight attendants on Tuesday, after a federal judge turned aside a union's bid to block the job losses.	American Airlines will recall hundreds of flight attendants as it steps up the number of flights it operates.	PP	False
4	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.	Cardinal Juan Jesus Posadas Ocampo died in 1993.	QA	True

Table 2.1: Examples of $T - H$ pairs in RTE-1 (Dagan *et al.*, 2005)

2.1.1.2 RTE-2

The RTE-2 dataset consists of 1600 $T - H$ pairs, divided into a development set and a test set, each containing 800 pairs. The texts T consist of 1 or 2 sentences, while the hypotheses H are usually made of a single short sentence. The focus here was on four out of the seven applications presented in RTE-1, namely IR, IE, QA, and multi-document summarization (equivalent to the CD task in RTE-1). Within each application setting the annotators selected 100 of both positive entailment examples, where T does entail H, as well as negative examples, where entailment does not hold for total of 200 pairs. Each pair was annotated with its related task (IE/IR/QA/SUM) and entailment judgment. A sample of the development set is given in Table 2.2. The examples in the dataset are based mostly on outputs (both correct and incorrect) of Web-based systems, while most of the input was sampled from existing application-specific benchmarks. Thus, the examples give some sense of how existing systems could benefit from an entailment engine postprocessing their output. This dataset was collected with regard to the following text processing applications:

- **Collecting IE pairs**: This task adapts the IE (and Relation Extraction) application settings to pairs of texts in contrast to text and a structured template. The pairs were generated using four different approaches. In the first approach, ACE-2004[1] relations were taken as templates for hypotheses. Relevant news articles were collected as texts T and then given to actual IE systems. The system outputs were used as hypotheses, generating both positive examples (from correct outputs) and negative examples (from incorrect outputs). In the second approach, the output of IE systems on a dataset of the MUC-4[2] was similarly used to create entailment pairs. the third subset consists of entailment pairs that were manually generated from the annotated MUC-4 dataset and news articles collected for the ACE relations. For example, given the ACE relation "X work for Y" and the text "An Afghan interpreter, employed by the United States, was also wounded.", a hypothesis "An interpreter worked for Afghanistan." is created, producing a non-entailing pair. In the forth approach, hypotheses which correspond to types of semantic relations not found in the ACE and MUC datasets were manually generated for sentences from the collected news articles. These relations were taken from various topic, such as sports, entertainment and science. All these processes simulate the need of IE systems to recognize that the given text indeed entails the semantic relation that is expected to hold between the candidate template slot fillers;

- **Collecting IR pairs**: In this application setting, the hypotheses are IR queries, which specify some statement, e.g. "Alzheimer's disease is treated using drugs". The hypotheses were adapted and simplified from standard IR evaluation datasets. For each hypothesis H several texts T that do or do not entail the hypothesis were selected from documents retrieved by different search engines (e.g. Google, Yahoo and MSN). In this application setting it is assumed that relevant documents should entail the given hypothesis;

- **Collecting QA pairs**: Annotators were given questions, taken from TRECQA[3] and QA@CLEF[4] datasets and the corresponding answers extracted from the Web by QA systems. Their task was to transform the question-answer pairs into text-hypothesis pairs following a two-stage routine: First, the annotators picked from the answer passage an answer term, either a correct or an incorrect one. Then, the annotators turned the question into an affirmative sentence including the answer term. These affirmative sentences serve as the hypotheses H, and the original answer passage serves as the text T. For example, given the question "How many inhabitants does Slovenia have?" and an answer text "In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers" T, the annotators picked "2 million inhabitants" as the correct answer term, which was used to turn the question into the statement "Slovenia

[1] ACE 2004 information extraction templates, from the NIST - `http://www.nist.gov/speech/tests/ace/` [Last access: 14th December, 2013]

[2] Message Understanding Conference, 1992

[3] `http://trec.nist.gov/data/qa.html` [Last access: 14th December, 2013]

[4] `http://clef-qa.fbk.eu/` [Last access: 14th December, 2013]

has 2 million inhabitants" H, producing a positive entailment pair;

- **Collecting SUM pairs**: In this setting T and H are sentences taken from a cluster of news documents, a collection of news articles that describe the same news topic. The annotators considered the output of a multi-document summarization systems, including the document clusters and the summary generated for each cluster. The annotators picked sentence pairs with high lexical overlap, preferably where at least one of the sentences was taken from the summary. Positive examples were constructed by simplifying the hypothesis by removing sentence parts, until it was fully entailed by T. Negative examples were simplified in the same manner. This simulates the summarization process of identification and removal of the redundant information from a text.

In a cross-annotation process of the collected pairs each pair was judged by at least two annotators. The average agreement on the test set, was 89.2%, which is an upper boundary of what could be expected from an entailment detection system. About 18% of the pairs were removed from the test set due to disagreement. The following situations often caused disagreement:

A number of reasons caused disagreement and bellow are listed some of the most permeative ones:

- T gives approximate numbers and H gives exact numbers;

- T states an asserted claim made by some entity, and the H drops the assertion and just states the claim. For example: T: "Scientists say that global warming is made worse by human beings.", H: "Global warming is made worse by human beings.";

- T makes a weak statement, and H makes a slightly stronger statement about the same thing.

Additional filtering was done which discarded pairs that seemed controversial, too difficult, or redundant. In this phase, about 256% of the original pairs were removed from the test set and minimal correction of texts was performed, e.g. fixing spelling and punctuation.

ID	TEXT	Hypothesis	TASK	JUDGMENT
77	Google and NASA announced a working agreement, Wednesday, that could result in the Internet giant building a complex of up to 1 million square feet on NASA-owned property, adjacent to Moffett Field, near Mountain View.	Google may build a campus on NASA property.	SUM	YES
110	Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."	A case of rabies was confirmed.	IR	NO
294	Meanwhile, in an exclusive interview with a TIME journalist, the first oneonone session given to a Western print publication since his election as president of Iran earlier this year, Ahmadinejad attacked the "threat" to bring the issue of Iran's nuclear activity to the UN Security Council by the US, France, Britain and Germany.	Ahmadinejad is a citizen of Iran.	IE	YES
387	About two weeks before the trial started, I was in Shapiro's office in Century City.	Shapiro works in Century City.	QA	YES
415	The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them.	Alzheimer's disease is treated using drugs.	IR	YES
691	Arabic, for example, is used densely across North Africa and from the Eastern Mediterranean to the Philippines, as the key language of the Arab world and the primary vehicle of Islam.	Arabic is the primary language of the Philippines.	QA	NO

Table 2.2: Examples o $T - H$ pair in RTE-2 (Ido *et al.*, 2006)

2.1.1.3 RTE-3

As in the previous challenges, the RTE-3 dataset consisted of 1600 text-hypothesis pairs, equally divided into a development set and a test set. While the length of the hypotheses H was the same as in the past datasets, a certain number of texts T were longer than in previous datasets.
The longer texts were marked as L, after being selected automatically when exceeding 270 bytes. In the test set they were about 17% of the total.
As in RTE-2, four applications – namely IE, IR, QA and SUM – were considered as settings or contexts for the pairs generation. 200 pairs were selected for each application in each dataset. Although the datasets were supposed to be perfectly balanced, the number of negative examples were slightly higher in both development and test sets (51.50% and 51.25% respectively; this was unintentional). Positive entailment examples, where T entailed H, were annotated YES; the negative ones, where entailment did not hold, NO. Each pair was annotated with its related task (IE/IR/QA/SUM) and entailment judgment (YES/NO, obviously released only in the development set). Table 2.3 shows some examples taken from the development set.
As in RTE-2, human annotators generated $T - H$ pairs within four application settings, following exactly the same process as used in RTE-2.
Each pair of the dataset was judged by three annotators. As in previous challenges, pairs on which the annotators disagreed were filtered-out. On the test set, the average agreement between each pair of annotators who shared at least 100 examples was 87.8%, with an average Kappa level of 0.75, regarded as substantial agreement according to Landis & Koch (1977).
19.2% of the pairs in the dataset were removed from the test set due to disagreement. The disagreement was generally due to the fact that the H was more specific than the T, for example because it contained more information, or made an absolute assertion where T proposed only a personal opinion. In addition, 9.4% of the remaining pairs were discarded, as they seemed controversial, too difficult, or too similar when compared to other pairs.
As far as the texts extracted from the web are concerned, spelling and punctuation errors were sometimes fixed by the annotators, but no major change was allowed, so that the language could be grammatically and stylistically imperfect. The hypotheses were finally double-checked by a native English speaker.

TASK	TEXT	HYPOTHESIS	ENTAILMENT
IE	At the same time the Italian digital rights group, Electronic Frontiers Italy, has asked the nation's government to investigate Sony over its use of anti-piracy software.	Italy's government investigates Sony.	NO
IE	Parviz Davudi was representing Iran at a meeting of the Shanghai Co-operation Organisation (SCO), the fledgling association that binds Russia, China and four former Soviet republics of central Asia together to fight terrorism	China is a member of SCO.	YES
IR	Between March and June, scientific observers say, up to 300,000 seals are killed. In Canada, seal-hunting means jobs, but opponents say it is vicious and endangers the species, also threatened by global warming.	Hunting endangers seal species.	YES
IR	The Italian parliament may approve a draft law allowing descendants of the exiled royal family to return home. The family was banished after the Second World War because of the King's collusion with the fascist regime, but moves were introduced this year to allow their return.	Italian royal family returns home.	NO
QA	Aeschylus is often called the father of Greek tragedy; he wrote the earliest complete plays which survive from ancient Greece. He is known to have written more than 90 plays, though only seven survive. The most famous of these are the trilogy known as Orestia. Also well-known are The Persians and Prometheus Bound.	"The Persians" was written by Aeschylus.	YES
SUM	A Pentagon committee and the congressionally chartered Iraq Study Group have been preparing reports for Bush, and Iran has asked the presidents of Iraq and Syria to meet in Tehran.	Bush will meet the presidents of Iraq and Syria in Theran.	NO

Table 2.3: Examples o $T - H$ pair in RTE-3 (Giampiccolo *et al.*, 2007)

2.1.1.4 RTE-4

In RTE-4, in participating systems were assigned the task of RTE in a set of 1000 $T - H$ pairs; i.e., they were required to decide, given a set of text pairs, called T and H, whether T entailed H or not.
Unlike the previous challenges, the main RTE-4 task asked the systems to make a three-way decision, further distinguishing, in case there was no entailment between T and H, whether the truth of H was contradicted by T, or remained unknown on the basis of the information contained in T.
In other words, the participating systems had to decided whether:

- T entailed H - in which case the pair was marked as **ENTAILMENT**;

- T contradicted H - in which case the pair was marked as **CONTRADICTION**;

- The truth of H could not be determined on the basis of T - in which case the pair was marked as **UNKNOWN**.

The classic two-way RTE task was also offered, in which the pairs where T entailed H were marked as ENTAILMENT, and those where the entailment did not hold were marked as NO ENTAILMENT. No development set was provided this challenges, as the pairs proposed were very similar to the ones contained in previous challenges development and test sets, which could therefore be used to train the systems.
Four application - namely IE, IR, QA and SUM - were considered as settings or contexts for the pairs generation. The length of the $H's$ was the same as in the past datasets; however, the T's were generally longer, following the decision taken last year of moving towards real cases where more discourse analysis is required. A major difference with respect to previous campaigns was that the RTE-4 dataset consisted of 1000 $T - H$ pairs, instead of 800. This was due to the fact that while 200 pairs were selected for QA and SUM, 300 were chosen for IE and IR, as these two settings proved somewhat more difficult in the previous challenges. The distribution according to the 3 way annotation, both in the individual settings and in the overall test set, was as follows: 50% ENTAILMENT, 35% UNKNOWN and 15% CONTRADICTION.

TASK	TEXT	HYPOTHESIS	ENTAILMENT
IE	Admiral Kuroyedov was in charge of the navy during the Kursk disaster of 2000, in which 118 sailors died when their submarine sank. Kuroyedov is being replaced by Vladimir Masorin, who was previously serving as the Chief of staff for the Russian Navy.	Kuroyedov caused the Kursk disaster	UNKNOWN
IE	Spencer Dryden, the drummer of the legendary American rock band Jefferson Airplane, passed away on Tuesday, Jan. 11. He was 66. Dryden suffered from stomach cancer and heart disease.	Spencer Dryden died at 66.	ENTAILMENT
IR	The Dalai Lama today called for Tibetans to end protests against the Beijing Olympics, also telling MPs in London he would happily accept an invitation to attend the event if relations with China improved.	China hosts Olympic games.	ENTAILMENT
IR	Lower food prices pushed the UK's inflation rate down to 1.1% in August, the lowest level since 1963. The headline rate of inflation fell to 1.1% in August, pushed down by falling food prices.	Food prices are on the increase.	CONTRADICTION
QA	The gambusia affinis, dubbed the mosquito fish, is an aquatic predator that devours mosquito larvae. Officials are releasing the fish into the fetid waters of abandoned pools to reduce the burgeoning mosquito population.	Gambusia is a species of mosquito	CONTRADICTION
QA	Four people were killed and at least 20 injured when a tornado tore through an Iowa boy scout camp on Wednesday, where dozens of scouts were gathered for a summer retreat, state officials said.	Four boy scouts were killey by a tornado.	UNKNOWN
SUM	Kingdom flag carrier British Airways (BA) has entered into merger talks with Spanish airline Iberia Lineas Aereas de Espana SA. BA is already Europe's third-largest airline.	The Spanish airline Iberia Lineas Aereas de Espana SA is Europ's third-largest airline.	CONTRADICTION

Table 2.4: Examples o $T - H$ pair in RTE-4 (Giampiccolo *et al.*, 2008)

As usual, human annotators generated $T - H$ pairs within the four aforementioned application settings, following exactly the same process as used in RTE-3.
As in previous challenges, each pair of the dataset was judged by three annotators. Pairs on which the annotators disagreed were discarded. The disagreement between annotators was often due to the fact that one annotator did not consider that some extra information was contained in the H, making it more specific than the T. In other cases, the disagreement was about whether the information in H was contradictory with respect to the content of T, or simply not sufficient to determine a judgment, especially in some ambiguous cases. Some pairs were also discarded because they were too similar to others, or their content was otherwise inappropriate.
Both texts and hypotheses were revised by native English speakers to eliminate the major spelling and grammar mistakes frequently present in texts taken from the web. No major changes were otherwise made, in order to keep the exercise realistic.

2.1.1.5 RTE-5

The RTE-5 was kept very similar to that proposed in RTE-4, in order to facilitate the comparison between the performances of systems which had participated in the previous challenges and encourage new participants to take part in an exercise not too different from last year's task. Nevertheless, some changes were introduced in order to move towards a more realistic exercise, stimulating researchers who had already participated in other RTE challenges to further test their systems against more challenging data sets.
First of all, while the length of the $H's$ was the same as in the past data sets (around 8 words), in the RTE-5 data set $T's$ were longer, up to 100 words, whereas in RTE-4 the average length was about 40 words. This length was meant to represent the average portion of the source document that a reader would naturally select, such as a paragraph or a group of related sentences. On the other hand, longer texts introduced in the exercise discourse phenomena, such as coreference, which were not present in the previous data sets. Moreover, texts, taken from a variety of freely available sources to avoid copyright problems, were not edited from their source documents. In this way, systems were asked to handle real text that may include typographical errors and ungrammatical sentences. For the rest, the basic structure of the challenge remained unchanged. Like in the previous RTE-4 challenge, both the classic two-way task and the three-way task were offered. In the traditional two-way task the pairs where T entails H are marked as ENTAILMENT, and those where the entailment does not hold are marked as NO ENTAILMENT.
The three-way task requires to further distinguish, in case there is no entailment between T and H, whether the truth of H is contradicted by T, or remains unknown on the basis of the information contained in T. In other words, the systems participating in the three-way task have to decide whether: T entails H - ENTAILMENT; T contradicts H - CONTRADICTION; The truth of H cannot be determined on the basis of T - UNKNOWN.

TASK	TEXT	HYPOTHESIS	ENTAILMENT
QA	The Grapes of Wrath, published exactly 70 years ago, can be seen as a prophetic novel, rooted in the tragedies of the Great Depression, but speaking directly to the harsh realities of 2009, writes Steinbeck scholar Robert DeMott. Steinbeck's epic novel, which traces harrowing exodus of Tom Joad and his family from blighted Oklahoma (where they are evicted from their farm), across the rugged American south-west via Highway 66, and on to what they mistakenly hope will be a more promising future in California, is considered by many readers to be the quintessential Depression-era story, and an ironic reversal of the rags-to-riches tale favoured by many optimistic Americans.	"The Grapes of Wrath" was written by Steinbeck.	ENTAILMENT
IR	Henan province has registered seven dead children and 4,761 HFMD cases. Shandong has reported five children dead from HFMD and 3,280 cases to deal with. HFMD can start from a variety of viruses of which Enterovirus 71 (EV-71) is the most common, followed by the Coxsackie A virus (Cox A16). There is an Incubation period from time of contact to appearance of symptoms between three to seven days.	Shandong is not far from Henan province.	UNKNOWN
IE	An appeals court in Eastern France has confirmed the Swedish car manufacturer Volvo is guilty over the deaths of two schoolchildren aged nine and ten and the serious injury of a third after a brakes failure caused an accident in 1999. The Volvo 850 TDI was being driven by a local teacher when it struck the children, who had been on their way to school. Driver Catherine Kohtz later asserted that the brake pedal had become stiff and the brakes themselves unresponsive as she traveled along the steep road.	Volvo is a car manufacturer from Finland.	CONTRADICTION

Table 2.5: Examples o $T - H$ pair in RTE-5 (Bentivogli *et al.*, 2009)

The settings from which the pairs were manually created by human annotators were IE, IR, QA. SUM was not considered in this challenge, as the Pilot Search data sets were entirely based on the Summarization setting. Table 2.5 presents some examples of $T - H$ pairs taken from the RTE-5 data set. The RTE-5 data set consisted of 1.200 $T - H$ pairs - 400 for each setting - equally divided into a Development Set and a Test Set. The distribution according to the 3 way annotation, both in the individual settings and in the overall data set, was 50% ENTAILMENT, 35% UNKNOWN, 15% CONTRADICTION.

As in the previous challenges, the overall process of data set creation requires the generation of large amounts of $T - H$ pairs, which are subsequently filtered to retain only those (i) featuring full agreement among three annotators in terms of the assigned entailment judgment, and (ii) compliant with the RTE guidelines for the creation of entailment pairs. The effort required to create the pairs varies a lot depending on the application scenario (being the QA pairs the most difficult to create and the IR pairs the easiest ones), and the type of entailment pair to be created (entailment, unknown, contradiction).

On average, six pairs per hour are created and annotated for the first time by an expert annotator. The subsequent entailment annotation of the existing pairs is much less time-consuming, as forty pairs per hour can be annotated.

As regards the RTE-5 data set, around 25% of the pairs originally created were discarded due to disagreement, and another 20% because they were unsuitable according to the guidelines (e.g. $T's$ too short or too long, ENTAILMENT pairs with the elements relevant to the entailment judgment repeated verbatim, or UNKNOWN pairs with T and H completely unrelated).

2.1.1.6 Summary

From its beginning in 2005, the task of RTE has evolved significantly, although its basic structure has been maintained in the years. In the first three challenges the task consisted of assigning a two-way entailment judgment (YES/NO) to a set of $T - H$ pairs. In RTE-4 and RTE-5 an additional 3-way judgment task was proposed together with the original one. In this task, in case of no entailment between T and H, systems have to specify whether T contradicts H (CONTRADICTION judgment), or the truth of H cannot be determined on the basis of T (UNKNOWN judgment).

In all the editions of the Challenge, the $T - H$ pairs were created by expert annotators from a number of NLP application settings. In the first challenge the applications considered were IR, CD, RC,QA, IE, MT and PP. In the following three challenges they were limited to IE, IR, QA, and SUM. In RTE-5 only IE, IR and QA were considered.

Table 2.6 shows how the composition of the data sets evolved over the years, in terms of number of pairs, T and H length, and word overlap between T and H. As far as the length of T and H is concerned, while $H's$ length remained constant over the years, the length of $T's$ substantially increased, passing from an average of 24.78 words in the RTE-1 Development set to around 100

words in the RTE-5 data sets. This gradual change to longer texts allowed for the introduction of discourse phenomena in the data set, which represented a first step towards the more realistic scenario proposed in the RTE-5 Search Pilot Task, where TE was performed against a real corpus.

Table 2.6 also shows data about the average word overlap between T and H, which is calculated counting all the words shared by T and H, and normalizing the results by the length of H. Overlap rates are grouped on the basis of the entailment judgment (YES/NO) assigned to the pairs. In general, it can be seen that positive examples (entailment=YES) show a higher word overlap with respect to the negative ones. Moreover, it is interesting to analyze the difference in word overlap between positive and negative pairs. This difference steadily increased over the years, reaching its highest value in the RTE-3 data sets, where the average overlap for positive pairs amounts to 71% whereas for negative pairs it amounts to 54%. This suggests that, for systems taking word overlap into account, the RTE-3 data set is potentially easier to process.

RTE-4 and RTE-5 data sets are different from the previous ones, due to the introduction of the three-way classification of the pairs. If we consider the class of NO-ENTAILMENT pairs, on the one hand we see a large difference in word overlap between UNKNOWN and ENTAILMENT pairs (similar to that present in the RTE-3 data set); on the other hand, CONTRADITION pairs present a high word overlap, very similar to that of ENTAILMENT pairs. This makes the RTE-4 and RTE-5 particularly challenging, as a part of the negative pairs are not distinguishable from the positive pairs by simply considering the word overlap feature.

Challenge	Data Set	Pairs	H length[1]	T length[2]	T/H Overlap (%)		
					YES	NO ENTAILMENT	
						UNKNOWN	CONTRADICTION
RTE-1	DEV	567	10.08	24.78	69.25	62.94	
	TEST	800	10.8	26.04	68.64	64.12	
RTE-2	DEV	800	9.65	27.15	69.1	58.16	
	TEST	800	8.39	28.37	70.63	63.32	
RTE-3	DEV	800	8.46	34.98	72.18	53.24	
	TEST	800	7.87	30.06	69.62	55.54	
RTE-4	TEST	1.000	7.7	40.15	68.95	57.36	67.97
RTE-5	DEV	600	7.79	99.49	77.71	61.95	77.06
	TEST	600	7.92	99.41	77.14	62.28	78.93

Table 2.6: RTE - 1 to RTE - 5 data sets

[1]average words in H
[2]average words in T

2.1.2 Relevant Resources and Tools

2.1.2.1 Evaluation Measures

The evaluation of all runs submitted was automatic, the judgments returned by the system being compared to the Gold Standard compiled by the human assessors. The main evaluation measure was accuracy, i.e., the fraction of correct answers. For the two-way task, a judgment of "NO ENTAILMENT" in a submitted run was considered to match either "CONTRADICTION" or "UNKNOWN" in the Gold Standard.

As a second measure, an Average Precision score was computed for systems that provided as output a confidence-ranked list of all test examples. Average Precision is a common evaluation measure for system rankings, and is computed as the average of the system's precision values at all points in the ranked list in which recall increases, that is at all points in the ranked list for which the gold standard annotation is ENTAILMENT. In other words, this measure evaluates the ability of systems to rank all the $T - H$ pairs in the test set according to their entailment confidence (in decreasing order from the most certain entailment to the least certain). More formally, it can be written as follows:

$$\frac{1}{R}\sum_{i=1}^{n}\frac{E(i) \times \sharp EntailmentUpTpPair(i)}{i} \tag{2.1}$$

where n is the number of the pairs in the test set, R is the total number of ENTAILMENT pairs in the Gold Standard, $E(i)$ is 1 if the $i - th$ pair is marked as ENTAILMENT in the Gold Standard and 0 otherwise, and i ranges over the pairs, ordered by their ranking.

In practice, the more confident the system was that T entailed H, the higher the ranking of the pair was. A perfect ranking would have placed all the positive pairs (for which the entailment holds) before all the negative ones, yielding an average precision value of 1. As average precision is relevant only for a binary annotation, in the case of three-way judgment submissions the pairs tagged as CONTRADICTION and UNKNOWN were conflated an retagged as NO ENTAILMENT.

2.1.2.2 First Challenge

In an overview of the systems participating in the first RTE challenge[1], of 2005, we saw that the main approaches (the best results), used are based on word overlap (Herrera *et al.*, 2005), statistical lexical relations (Bayer *et al.*, 2005), WordNet (Miller, 1995)[2] similarities (Herrera *et al.*, 2005), syntactic matching (Delmonte *et al.*, 2005), world knowledge (Bayer *et al.*, 2005), edit distance between parsing trees (Kouylekov & Magnini, 2005). The majority of the systems experiment with different threshold and parameter settings to estimate the best performance. The parameter adjustment process is related to the carrying out of numerous experiments and still the settings selected after these experiments may lead to incorrect reasoning.

[1] `http://pascallin.ecs.soton.ac.uk/Challenges/RTE/` [Last access: 14th December, 2013]

[2] `http://wordnet.princeton.edu/` [Last access: 14th December, 2013]

Participants	*Accuracy*	System Description				
		Word Overlap	Statistical Lexical Relations	WordNet	Syntactic Matching	Logical Inference
Delmonte *et al.* (2005)	0,606			X	X	X
Bayer *et al.* (2005)	0,586		X			
Glickman & Dagan (2005)	0,586		X			
Herrera *et al.* (2005)	0,566	X	X		X	
Kouylekov & Magnini (2005)	0,559	X				

Table 2.7: Best Results in RTE-1 (Dagan *et al.*, 2005)

In Delmonte *et al.* (2005), the system for semantic evaluation VENSES (Venice Semantic Evaluation System) is organized as a pipeline of two subsystems: the first is a reduced version of *GETARUN*, our system for Text Understanding. The output of the system is a flat list of head-dependent structures with Grammatical Relations and Semantic Roles labels. The evaluation system is made up of two main modules: the first is a sequence of linguistic rule-based subcalls; the second is a quantitatively based measurement of input structures. VENSES measures semantic similarity which may range from identical linguistic items, to synonymous or just morphologically derivable. Both modules go through General Consistency checks which are targeted to high level semantic attributes like presence of modality, negation, and opacity operators, temporal and spatial location checks.

Bayer *et al.* (2005) intended to exemplify two different ends of the spectrum of possibilities. The first submission is a traditional system based on linguistic analysis and inference, while the second is inspired by alignment approaches from MT.

In Glickman & Dagan (2005) proposes a general probabilistic setting that formalizes the notion of TE and describe a model for lexical entailment based on web co-0ccurrence statistic in a bag of words representation.

The system described in Herrera *et al.* (2005), is based on the use of a broad-coverage parser to extract dependency relations and a module which obtains lexical entailment relations from WordNetMiller (1995).

The transformation-based entailment method makes use of various types of entailment knowledge to gradually transform T such that it becomes more similar to H, or vice versa. Kouylekov & Magnini (2005) assumed a distance-based framework,where the distance between T and H is

inversely proportional to the entailment relation in the pair, estimated as the sum of the costs of the edit operations (i.e. insertion, deletion, substitution) on the parse tree, which are necessary to transform T into H. They use different resources to estimate the edit operations cost and to ensure the non-symmetric directionality of the entailment relation.

2.1.2.3 Second Challenge

In the second edition, of 2006, the main directions were generally the same, only algorithms were more sophisticatedly and also the results were better (average precision grew up from 55.12 % in 2005 to 58.62 % in 2006). New directions are related to semantic role labelling (Hickl *et al.*, 2006), Machine Learning classification, using of background knowledge (Tatu *et al.*, 2006) acquisition of entailment corpora (Hickl *et al.*, 2006). Some groups tried to detect non entailment, by looking for various kinds of mismatch between the text and the hypothesis.

Participants	***Accuracy***	System Description								
		Lexical Relation DB	n-gram / Subsequence Overlap	Syntactic Matching / Alignment	Semantic Role Labelling / Framenet / Propbank	Logical Inference	Corpus / Web-Based Statistics	ML Classification	Paraphrase Templates / Background Knowledge	Acquisition of Entailment Corpora
Hickl *et al.* (2006)	0,754	X	X	X	X		X	X		X
Tatu *et al.* (2006)	0,738	X				X			X	
Zanzotto *et al.* (2006)	0,639	X		X			X	X		
Adams *et al.* (2006)	0,626	X						X		
Bos & Markert (2006)	0,616	X					X	X		

Table 2.8: Best Results in RTE-2 (Ido *et al.*, 2006)

In Hickl *et al.* (2006), they introduce a new system for RTE (known as GROUNDHOG) which utilizes a

classification-based approach to combine lexico-semantic information derived from text processing applications with a large collection of paraphrases acquired automatically from the glswww. Trained on 200,000 examples of TE extracted from newswire corpora.

Logic inference can be considered as one of the most direct approaches to the entailment problem. Tatu *et al.* (2006) transformed two text snippets into three-layered semantically-rich logic form representations, generates an abundant set of lexical, syntactic, semantic, and world knowledge axioms and, iteratively, searches for a proof for the entailment between the text T and a possibly relaxed version of the hypothesis H. They could improve the performance of their system using the lexical inference system in combination with their logical approach.

The system described in Zanzotto *et al.* (2006) defines a cross-pair similarity measure based on the syntactic trees of T and H, and combines such similarity with traditional intra-pair similarities to define a novel semantic kernel function. The intuition behind this approach is that not only intra-pair similarity between T and H, but also cross-pair similarity between two pairs can be useful to address the problem. The latter similarity measure along with a set of annotated examples is used by a learning algorithm to automatically derive syntactic and lexical rules to solve complex entailment cases.

Adams *et al.* (2006) presents a system of TE based primarily on the concept of lexical overlap. The system begins with a bag of words similarity overlap measure, derived from a combination of WordNetMiller (1995) lexical chains to form a mapping of terms in the hypothesis to the source text. It then looks for negations not found in the mapping, and for the lexical edit distance of the mapping. These items are then entered into a decision tree to determine the overall entailment.

One of the first efforts to combine shallow NLP methods with a deep semantic analysis was made by Bos (2005). In RTE-2, Bos & Markert (2006) combined two approaches, a shallow method based mainly on word-overlap and a method based on logical inference, using first-order theorem proving and model building techniques. They used a machine learning technique to combine features from both methods.

2.1.2.4 Third Challenge

In the third edition, of 2007, we can notice a move toward deep approaches. The groups were oriented on the approaches based on the syntactic structure of Text and Hypothesis, on semantic understanding of the texts and also on verification of the content and new situations and contexts that meet in the test data. A special attention was given to the named entities, where (Tatu & Moldovan, 2007) had special rules for Person names, and where (Iftene & Balahur-Dobrescu, 2007) had special rules for all named entities. Some form of relation extraction has been introduced: information extracted automatically by a system (Hickl & Bensley, 2007). Also, in comparison to previous editions, now the longer texts need anaphora resolution (Iftene & Balahur-Dobrescu, 2007).

Participants	*Accuracy*	System Description								
		Lexical Relation, WordNet	n-gram similarity	Syntactic Matching / Alignment	Semantic Role Labelling / Framenet / Propbank, Verbnet	Logical Inference	Corpus / Web-Based Statistics, LSA	ML Classification	Anaphora resolution	Entailment Corpora - DIRT Background Knowledge
Hickl & Bensley (2007)	0,800	X	X			X		X	X	X
Tatu & Moldovan (2007)	0,723	X				X			X	X
Iftene & Balahur-Dobrescu (2007)	0,691	X		X						X
Adams *et al.* (2007)	0,670	X	X				X	X		
Wang & Neumann (2007)	0,669				X				X	

Table 2.9: Best Results in RTE-3 (Giampiccolo *et al.*, 2007)

The GROUNDHOG system (Hickl & Bensley, 2007) uses a pipeline of lightweight, largely statistical systems for commitment extraction, lexical alignment, and entailment classification in order to estimate the likelihood that T includes sufficient linguistic content to textually entail H.

As in RTE-2, Tatu & Moldovan (2007) based in logical inference, first, they summarize our semantic logical-based approach. The novelties include new resources, such as eXtended WordNet[1] (Harabagiu *et al.*, 1999) which provides a large number of world knowledge axioms, event and temporal information provided by the Temporal Awareness and Reasoning Systems for Question Interpretation (TARSQI)[2] toolkit, logic form representations of events, negation, coreference and context, and new improve-

[1] `http://xwn.hlt.utdallas.edu/` [Last access: 14th December, 2013]

[2] `http://www.timeml.org/site/tarsqi/` [Last access: 14th December, 2013]

ments of lexical chain axiom generation.

In order to boost the similarity scores and extend them to a different level, Iftene & Balahur-Dobrescu (2007) compared $H's$ parse tree against subtrees of $T's$ parse tree. They transformed the hypothesis making use of an extensive semantic knowledge from sources like Discovery of Inference Rules from Text (DIRT) (Lin & Pantel, 2001), WordNet, Wikipedia[1] and a database of acronyms. Additionally, they took advantage of hand coded complex grammar rules for rephrasing in English.

In Adams *et al.* (2007), two textual entailment approaches are presented. The first one is based primarily on the concept of lexical overlap, considering a bag-of-words similarity overlap measure to form a mapping of terms in the hypothesis to the source text. The second system is a lexico-semantic matching between the text and the hypothesis that attempts an alignment between chunks in the hypothesis and chunks in the text, and a representation of the text and hypothesis as two dependency graphs. Both approaches employ decision trees as a supervised learning algorithm.

The system presented in Wang & Neumann (2007) has moved from a puristic syntactic approach, in the sense that they only performed dependency parser, to the development of specialized RTE-modules capable of tackling more entailment phenomena. They present a novel approach to RTE that exploits a structure-oriented sentence representation followed by a similarity function. The structural features are automatically acquired from tree skeletons that are extracted and generalized from dependency trees. Their method makes use of a limited size of training data without any external knowledge bases or handcrafted inference rules. For preprocessing, they use a PoS tagger, a dependency parser[2], and a Named Entity (NE) recognizer[3] in order to annotate the original plain texts. The Precision-Oriented (PO) modules are created to specialize the system in the RTE task.

2.1.2.5 Fourth Challenge

Inside the different approaches to TE, the use of ML approaches is dominant. This is mainly because both logic and rule-based methods suffer from either limited coverage of hand-crafted rules and lower performance. In ML approaches, a variety of features including lexical, syntactic and semantic features can be extracted from training examples, thus can be employed to train a classifier, Bensley & Hickl (2008) focused on collecting deeper semantic features, in using a pipeline of lightweight, largely statistical systems for commitment extraction, lexical alignment, and entailment classification in order to estimate the likelihood that a T includes the linguistic content sufficient to textually entail a H.

The main idea in Iftene (2008) is to map every word from hypothesis to one or more words from the text. For that, this system transform the hypothesis making use of extensive semantic knowledge from sources like DIRT, WordNetMiller (1995), VerbOcean[4], Wikipedia and a database of acronyms.

[1] `http://en.wikipedia.org/` [Last access: 14th December, 2013]

[2] The dependency parser used is MINIPAR (Lin, 1998)

[3] They used the Stanford NE recognition system (Finkel *et al.*, 2005)

[4] `http://demo.patrickpantel.com/demos/verbocean/` [Last access: 14th December, 2013]

After the mapping process, they associate a local fitness value to every word from hypothesis, which is used to calculate a global fitness value for current fragments of text. The global fitness value is decreased in cases in which a word from hypothesis cannot be map to one word from the text or when we have different forms of negations for mapped verbs. In the end, using thresholds identified in the training step for global fitness values, they decide for every pair from test data if they have entailment or not.

Participants	***Accuracy***	System Description								
		Lexical Relation, WordNet	n-gram similarity	Syntactic Matching / Alignment	Semantic Role Labelling / Framenet / Propbank, Verbnet	Logical Inference	Corpus / Web-Based Statistics, LSA	ML Classification	Anaphora resolution	Entailment Corpora - DIRT Background Knowledge
Bensley & Hickl (2008)	0,746	X	X			X		X	X	X
Iftene (2008)	0,721	X		X						X
Wang & Neumann (2008)	0,706			X				X		
Li *et al.* (2008)	0,659	X	X					X		
Balahur *et al.* (2008)	0,608	X	X	X						

Table 2.10: Best Results in RTE-4 (Giampiccolo *et al.*, 2008)

The approach proposed in Wang & Neumann (2008) is based on constructing structural features from the abstract tree descriptions, which are automatically extracted from syntactic dependency trees of T and H. These features are then applied by a subsequence-kernel-based classifier that learns to decide whether the entailment relation holds between two texts. A divide-and-conquer architecture is then in charge of providing a set of specific RTE methods (namely: temporal anchors, named entities and noun phrase anchors), and then combine them applying a voting scheme in order to maximize the accuracy. In Li *et al.* (2008), they design different strategies to recognize true entailment and false entailment. The similarity between hypothesis and text is measured to recognize true entailment. They detect the exact entity and relation mismatch to recognize the false entailment.

The RTE system presented in Balahur *et al.* (2008) tackles the entailment phenomenon from two different points of view. First, they build the system's core by means of several lexical measures and further on, they add some semantic constraints that they think are appropriated for the entailment recognition. The reason for creating this core was given by (i) the fact that the integration of more complex semantic knowledge is a delicate task and it would be easier if they had a solid base system; and (ii) although the proposed core needs some language dependent tools (e.g. lemmatizer, stemmer), it could be easily ported to other languages.

2.1.2.6 Fifth Challenge

In this challenge the best result for two way is Iftene & Moruz (2009), the main idea of they system is to map every word in the hypothesis to one or more words in the text. For that, they transform the hypothesis, using extensive semantic knowledge from sources (as previous RTE). The main improvement this challenge was related to the pre-processing part, the texts were obtained from a variety of sources and were not edited from their source documents, they focused on this part. Thus, they identify and eliminate special characters that occur frequently on web pages. This choice is based on the fact that with or without these characters the meaning of the text is the same, but the quality of the tools output is improved. Additionally, they process the LingPipe[1] output with GATE Cunningham *et al.* (2002) in order to identify some named entities categories unidentified by LingPipe such as nationality, language, and job.

In Wang *et al.* (2009) propose a joint syntactic-semantic representation to better capture the key information shared by the pair, and also apply a co-reference resolver to group cross-sentential mentionings of the same entities together.

In Li *et al.* (2009), they propose an interesting method, SEGraph (Semantic Elements based Graph). This method divides the Hypothesis and Text into two types of semantic elements: Entity Semantic Element and Relation Semantic Element. The SEGraph is then constructed, with Entity Elements as nodes, and Relation Elements as edges for both Text and Hypothesis. They recognize the textual entailment based on the SEGraph of Text and SEGraph of Hypothesis.

In Mehdad *et al.* (2009) use of semantic knowledge based on Wikipedia. More specifically, they used it to enrich the similarity measure between pairs of text and hypothesis (i.e. the tree kernel for text and hypothesis pairs), with a lexical similarity (i.e. the similarity between the leaves of the trees.

Sammons *et al.* (2009) present an approach to textual entailment recognition, in which inference is based on a shallow semantic representation of relations (predicates and their arguments) in the text and hypothesis of the entailment pair, and in which specialized knowledge is encapsulated in modular components with very simple interfaces. They propose an architecture designed to integrate different, unscaled NLP resources, and demonstrate an alignment-based method for combining them. They clarify the purpose of alignment in the RTE task, identifying two distinct

[1] `http://www.alias-i.com/lingpipe/` [Last access: 14th December, 2013]

alignment models, each of which leads to a different type of entailment system. They identify desirable properties of alignment, and use this to inform our implementation of an alignment component.

Participants	*Accuracy*	System Description								
		Lexical Relation, WordNet	n-gram similarity	Syntactic Matching / Alignment	Semantic Role Labelling / Framenet / Propbank, Verbnet	Logical Inference	Corpus / Web-Based Statistics, LSA	ML Classification	Anaphora resolution	Entailment Corpora - DIRT Background Knowledge
Iftene & Moruz (2009)	0,735	X	X			X		X	X	X
Wang *et al.* (2009)	0,685	X		X						X
Li *et al.* (2009)	0,670			X				X		
Mehdad *et al.* (2009)	0,662	X	X					X		
Sammons *et al.* (2009)	0,643	X	X	X						

Table 2.11: Best Results in RTE-5

2.1.2.7 Summary

The RTE systems results demonstrate general improvement with time, with overall accuracy levels ranging from 50% to 65% on RTE-1 (17 submissions), from 53% to 75% on RTE-2 (23 submissions), from 49% to 80% on RTE-3 (26 submissions), from 45% to 74% on RTE-4 (26 submissions, three-way task) and from 43% to 75% on RTE-5 (20 submissions, three-way task). Common approaches used by the submitted systems include ML, logical inference , cross-pair similarity measures between T and H and word alignment.

Challenge	Accuracy Average
RTE-1	0.581
RTE-2	0.675
RTE-3	0.711
RTE-4	0.688
RTE-5	0.679

Table 2.12: Average the top five results

2.2 Unsupervised Language-Independent Methodologies for RTE

Different approaches have been proposed to recognize Textual Entailment: from unsupervised language-independent methodologies Glickman & Dagan (2005), Perez *et al.* (2005) and Bayer *et al.* (2005) to deep linguistic analysis. We will particularly detail the unsupervised language-independent approaches, to which our work can be directly compared, at least to a certain extent. One of the most simple proposals (Perez *et al.*, 2005) explores the *BLEU algorithm* Papineni *et al.* (2002).

First, for several values of n (typically from 1 to 4), they calculate the percentage of n-grams from the text T, which appear in the hypothesis H. The frequency of each n-gram is limited to the maximum frequency with which it appears in any text T. Then, they combine the marks obtained for each value of n, as a weighted linear average and finally apply a brevity factor to penalize short texts T. The output of BLEU is then taken as the confidence score. Finally, they perform an optimization procedure to choose the best threshold according to the percentage of success of correctly recognized entailments. The value obtained was 0.157. Thus, if the BLEU output is higher than 0.157, the entailment is marked as true, otherwise as false. This procedure achieves $0,495$ of accuracy in recognizing TE.

In Bayer *et al.* (2005) the entailment data is treated as an aligned translation corpus. In particular, they use the *GIZA++* toolkit (Och & Ney, 2003) to induce alignment models. However, the alignment scores alone were next to useless for the RTE-1 development data, predicting entailment correctly only slightly above chance. As a consequence, they introduced a combination of metrics intended to measure translation quality. Finally, they combined all the alignment information and string metrics with the classical $K - NN$ classifier to choose, for each test pair, the dominant truth value among the five nearest neighbors in the development set. This method achieves $0,586$ of accuracy.

The most interesting work is certainly the one described in Glickman & Dagan (2005) who propose a general probabilistic setting that formalizes the notion of TE. Here, they focus on identifying when the lexical elements of a textual hypothesis H are inferred from a given text T. The probability of lexical entailment is derived from Equation 2.2 where $hits(.,.)$ is a function that returns the number of documents, which contain its arguments.

$$P(H|T) = \prod_{u \in H} max_{v \in T} \frac{hits(u,v)}{hits(v)} \tag{2.2}$$

The text and hypothesis of all pairs in the development and test sets were tokenized and stop words were removed to empirically tune a decision threshold, λ. Thus, for a pair $T - H$, they tagged an example as true (i.e. entailment holds) if $P(H|T) > \lambda$, and as false otherwise. The threshold was empirically set to 0.005. With this method accuracy of $0,586$ is achieved. The best results from these three approaches are obtained by Glickman & Dagan (2005), who introduce the notion of asymmetry within their model without clearly mentioning it. The underlying idea is based on the fact that for each word in H, the best asymmetrically co-occurring word in T is chosen to evaluate $P(H|T)$. Although all three approaches show interesting properties, they all depend on tuned thresholds, which can not reliably be reproduced and need to be changed for each new application. Moreover, they need training data, which may not be available. Our idea aims at generalizing the hypothesis made by Glickman & Dagan (2005). Indeed, their methodology is only based on one pair $(u,v), \forall u$ and does not take into account the fact that many pairs i.e. $(u,v), \exists v \forall u$ may help the decision process. Moreover, they do not propose a solution for the case where the ratio $\frac{hits(u,v)}{hits(v)}$ is null. Finally, we propose to avoid the definition of a "hard" threshold and study exhaustively asymmetry in language i.e. not just by the conditional probability as done in Glickman & Dagan (2005). For that purpose, we propose a new IAM called the AIS combined with different Association Measures.

CHAPTER 3

CORPUS CONSTRUCTION

"Men often become what they believe themselves to be.
If I believe I cannot do something, it makes me incapable of doing it.
But when I believe I can, then I acquire the ability to do it even if
I didn't have it in the beginning."
Mahatma Gandhi

In this chapter we detail our methodology for building a specialized corpus needed to evaluate our approach to identify entailment by generality in pairs of sentences. For this task, we use the technique of crowdsourcing, through the CrowdFlower service, which appears to be an excellent medium for large scale manual annotation.

3.1 Crowdsourcing

Large scale annotation projects such as TreeBank (Marcus *et al.*, 1993), PropBank (Palmer *et al.*, 2005), TimeBank (Pustejovsky *et al.*, 2003), FrameNet (Baker *et al.*, 1998), SemCor (Miller *et al.*, 1993), and others play an important role in NLP research, encouraging the development of novel ideas, tasks, and algorithms. The construction of these datasets, however, is extremely expensive in both annotator-hours and financial cost. Since the performance of many NLP tasks is limited by the amount and quality of data available to them (Banko & Brill, 2001), one promising alternative for some tasks is the collection of non-expert annotations.

The availability and the increasing popularity of crowdsourcing services have been considered as an interesting opportunity to meet the aforementioned needs and design criteria.

One of the most popular crowdsourcing services is MTurk[1], *"[…] a crowdsourcing Internet marketplace that enables computer programmers (known as Requesters) to co-ordinate the use of human intelligence to perform tasks which computers are unable to do […] The Requesters are able to pose tasks known as HITs (Human Intelligence Tasks) […] Workers [also known as "*Turkers*"] can then browse among existing tasks and complete them for a monetary payment set by the Requester. To place HITs, the requesting programs use an open Application Programming Interface […] Requesters*

[1] `http://www.mtruk.com/` [Last access: 14th December, 2013]

can ask that Workers fulfill Qualifications before engaging a task, and they can set up a test in order to verify the Qualification. They can also accept or reject the result sent by the Worker, which reflects on the Worker's reputation"[1].

Crowdsourcing services have been recently used with success for a variety of NLP applications (Callison-Burch & Dredze, 2010). Although MTurk is directly accessible only to US citizens, the CrowdFlower service[2] provides a crowdsourcing interface to MTurk for non-US citizens.

The main idea in using crowdsourcing to create NLP resources is that the acquisition and annotation of large datasets, needed to train and evaluate NLP tools and applications, can be carried out in a cost-effective manner by defining simple Human Intelligence Tasks (HITs) routed to a crowd of non-expert workers, called "*Turkers*", hired through on-line marketplaces. Figure 3.1 illustrates the MTurk process[3].

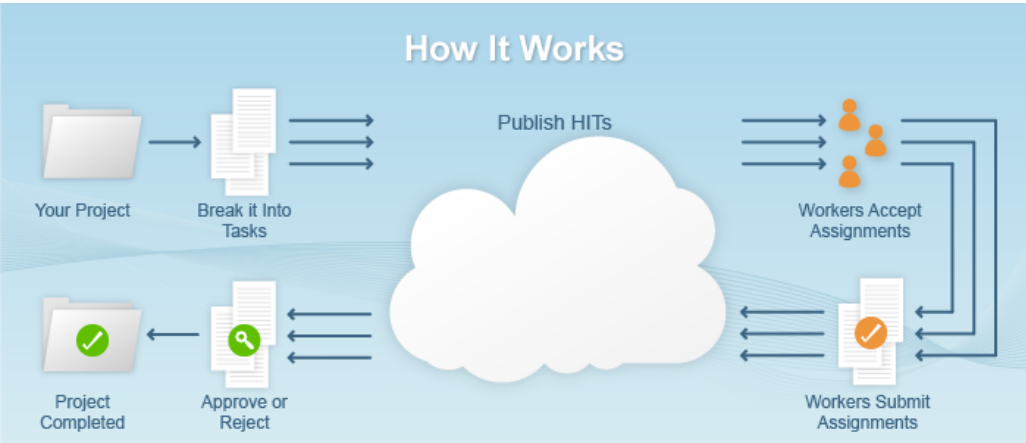

Figure 3.1: Mechanical Turk process.

3.2 Quality Control of Crowdsourced Data

The design of a data acquisition HITs has to take into account several factors, each having a considerable impact on the difficulty of instructing the workers, the quality and quantity of the collected data, the time and overall costs of the acquisition. In our particular case, *Turkers* are presented with the task of labeling input data referring to a fixed set of possible values (i.e. making a choice between multiple alternatives).

For annotation jobs, quality control mechanisms can be easily set up by calculating *Turkers* agreement, by applying voting schemes, or by adding hidden gold units (or test questions) to the data to be annotated. CrowdFlower provides means to check workers reliability, and weed out untrusted ones without money waste. This is achieved by adding hidden gold standard units in the

[1] Taken from `http://en.wikipedia.org/wiki/Amazon_Mechanical_Turk` [Last access: 14th December, 2013]

[2] `http://crowdflower.com/` [Last access: 14th December, 2013]

[3] Source: `https://requestersandbox.mturk.com/tour/how_it_works` [Last access: 14th December, 2013]

data to be annotated. For that purpose, we annotated 260 pairs that constitute our gold units[1].
As regards textual entailment, the first work exploring the use of crowdsourcing services for data annotation is Snow *et al.* (2008), which shows high agreement between non-expert annotations of the RTE-1 dataset and existing gold standard labels assigned by expert annotators.

3.3 Building Methodology

Our approach builds on a pipeline of HITs routed to MTurk workforce through the CrowdFlower interface. The objective is to collect $T - H$ pairs where entailment by generality holds.
Our building methodology has several stages, first we select the positive pairs of TE from the first five RTE challenges. These pairs are then submitted to CrowdFlower through a job that we have built online (see Appendix C), to be evaluated by *"Turkers"*[2]. In CrowdFlower each $T - H$ pair is a unit. The *"Turkers"* are asked to select either one of *"Entailment by Generality"*, *"Entailment, but not Generality"* or *"Other"* whichever is most appropriate for the $T - H$ pair under consideration (see Appendix C).
In the end we built our form, we calibrate and parametrize our job (sample, see figure 3.2). We define the following parameters[3]:

- ***"Gold Units"***
- ***"Data Settings"***
  - ***Make Your Data Public*** *- "This data may be used by the public for research purposes. By checking this box, you agree to the Terms of Service."*
- ***"Skill Requirements"***
  - ***Bronze*** *- "Bronze contributors are trusted contributors within our system. There are approximately 20,000 contributors in this group. Among other qualifying criteria, all Bronze contributors have seen at least 100 Gold units, and have achieved at least 80% accuracy on these units."*
- ***"Task Settings"***
  - ***Judgments per unit*** *- "This is the number of trusted judgments we will collect for each of your units.Gold units will receive a higher number of judgments."*
  - ***Units per page*** *- "This is the number of units that contributors will see on each page."*

[1]Our Gold Units are available at `http://hultig.di.ubi.pt/~sebastiao/` [Last access: 21th December, 2013]
[2]This dataset is available at `http://hultig.di.ubi.pt/~sebastiao/` [Last access: 21th December, 2013]
[3]Source: `https://crowdflower.com/resource-library` [Last access: 14th December, 2013]

- *"Contributor Pay"*

  - ***Seconds per unit*** *- "This is the number of seconds you think it will take the average contributor to complete a single unit on a page. This is only used to estimate how much contributors can earn per hour on this task. (They will not see this estimate.)"*

  - ***Payment per page (cents)***

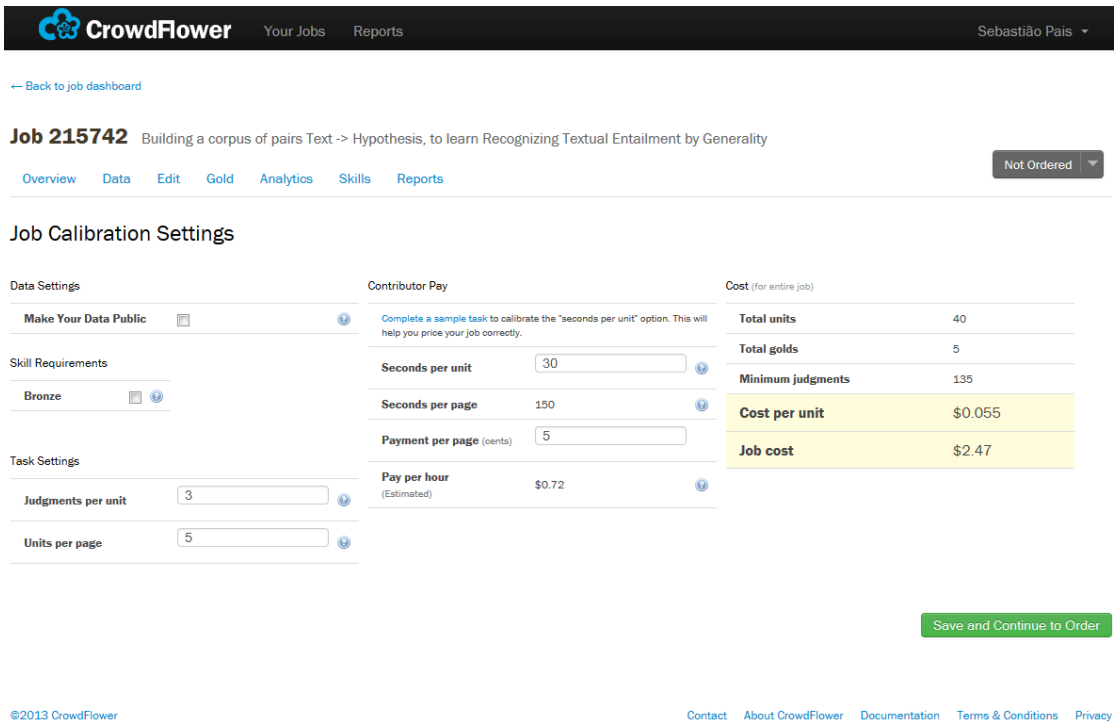

Figure 3.2: Job Calibration Settings in CrowdFlower.

Once calibrated and parameterized, we add money to our account of CrowdFlower, to submit our job for evaluation.

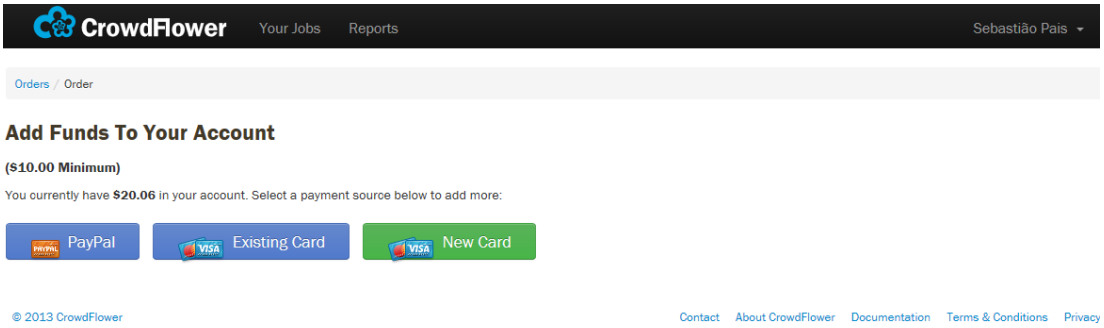

Figure 3.3: Add Funds in CrowdFlower.

When the *Turkers* provide the required number of trusted judgments and therefore finish the job, there is available a list of reports, Figure 3.4, to be further analyzed.

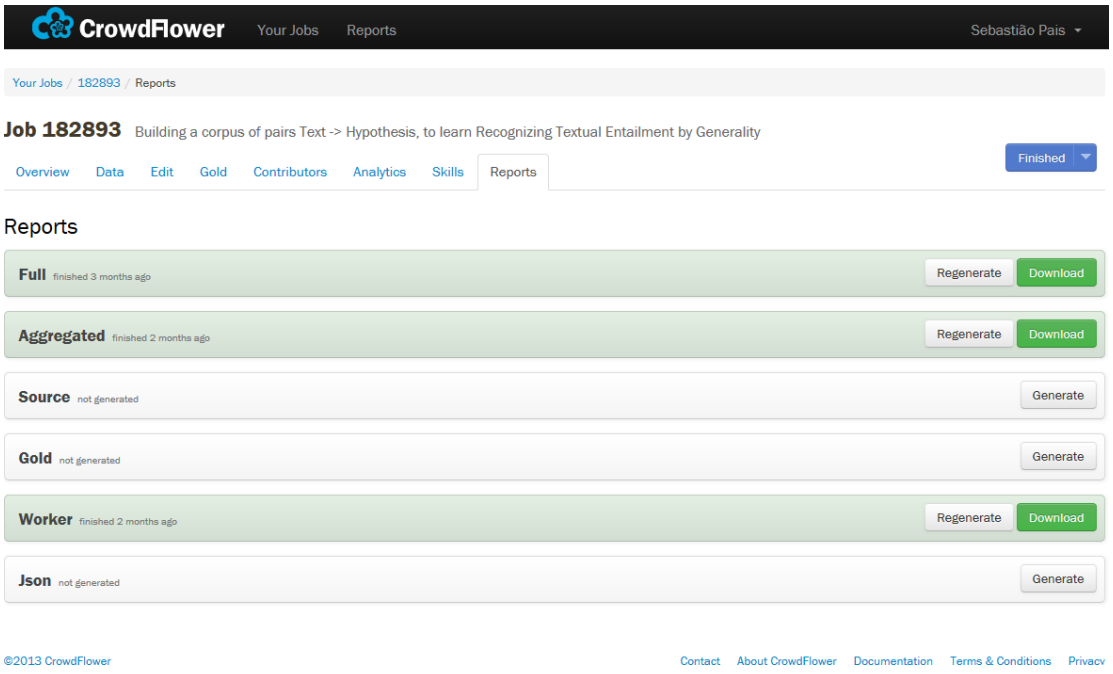

Figure 3.4: All Reports in CrowdFlower.

The ***Full*** report contains all the answers for all *Turkers*, i.e., we know the *Turkers* which responded and what their answers were.

The ***Aggregated*** report, is what we really want, i.e., it shows us what classification for each $T - H$ pair was the most favored one, where each pair can be classified as either *"Entailment by Generality"*, *"Entailment, but not Generality"* or *"Other"* and respectively level of inter-annotator agreement. It is this report from which we build our corpus of *"Entailment by Generality"* $T - H$ pairs.

3.4 Quantitative Analysis

Table 3.1 summarizes the work involved in the annotation of the entailment cases of **RTE-1** through **RTE-5** datasets with *"Entailment by Generality"*, *"Entailment, but not Generality"* and *"Other"* labels. We uploaded 2000[1] $T - H$ pairs known to be in entailment relation. Of those, 1740 were submitted for evaluation and 260 were *"Gold"* pairs.

[1]Input Pairs are available at `http://hultig.di.ubi.pt/~sebastiao/` [Last access: 21th December, 2013]

Table 3.1: Summary of RTE by Generality corpus annotation task

	RTE-1	RTE-2	RTE-3	RTE-4	RTE-5
# Input Pairs[1]	400	400	400	500	300
# Pairs to Launch[2]			1740		
# Gold Pairs[3]			260		
# Output Pairs[4]			1203		
# Discarded Pairs[5]			797		
Evaluation Time[6]			≈43 days		
# Trusted "Turkers"			2308		
# Trusted Judgments			5220 (1740*3)		
# Untrusted Judgments			60482		
Cost ($)			108,08		

In total, 2000 pairs were upload to CrowdFlower. Table 3.1 evidences that the annotated corpus contains 1203 $T - H$ pairs that are in TE by Generality relation.[6] Each case was evaluated by three *"Turkers"* for final average inter-annotator agreement of $0,8$.

This task proved to be complicated for the *"Turkers"*, as it is difficult for a human annotators to identify entailment relation and entailment by generality in particular. Proof of this is the time spent to complete the task (*Evaluation Time*) and the total number of *Judgments* (*Trusted* + *Untrusted*) needed to achieve the final objective.

The result of our work is the first large-scale dataset containing a reasonable number of pairs that are in TE by Generality relation.

[0]Number of pairs $T->H$ uploaded

[1]Number of pairs $T->H$ submitted for evaluation

[2]Number of *"Gold"* pairs $T->H$

[3]Number of pairs $T->H$ classified as *"Entailment by Generality"*

[4]Number of pairs $T->H$ classified as *"Entailment, but not Generality"* or *"Other"*

[5]Time accomplish the task

[6]The subset of 1203 TE by Generality pairs is available at `http://hultig.di.ubi.pt/~sebastiao/` [Last access: 21th December, 2013]

CHAPTER 4

OUR METHODOLOGY FOR RTE BY GENERALITY

"What ever the mind of man can conceive and believe,
it can achieve. "
Napoleon Hill

This chapter describes our approach to the problem of recognizing textual entailment by generality. In this, we treat a sentence as a string of lexical units or tokens. It turns out that there are different ways to tokenize and below we describe few of them. We will see in the following chapter that they possess interesting properties in the context of this work.

4.1 Contextual Word Similarity

A prerequisite of any language-independent NLP methodology is the capability to extract implicit and explicit knowledge from raw natural language texts as the basic textual information.
Two different types of knowledge can be acquired depending on the basic textual unit under study. On the one hand, analyzing word similarities evidences intrinsic knowledge about the language (i.e. information about the language which is not explicitly encoded in texts). Traditional examples are collocations and word semantic relations such as hypernymy/hyponymy, meronymy/holonymy, synonymy or antonymy, which must be mined from texts. On the other hand, explicit knowledge about the language (i.e. information about the message conveyed by the texts) can be extracted from the evaluation of sentence, passage and text similarities[1]. There are obviously some exceptions. In particular, analyzing sentence similarities in the context of topic segmentation is likely to identify intrinsic knowledge about discourse structure (Dias *et al.*, 2007).
Identifying different types of similarities between words has been an important goal in NLP. Usually it is achieved through some statistical approach for computing the degree of similarity between unit representation in an appropriate feature space. In this approach a word is represented by a word co-occurrence vector in which each entry corresponds to another word in the lexicon. The value of an entry specifies the frequency of joint occurrence of the two words in the corpus, that is, the

[1]From now on, we will refer to sentences, passages and texts simply as texts.

frequency with which they co-occur within some particular relationships in the text. The degree of similarity between a pair of words is then computed by some similarity or distance measure that is applied to the corresponding pair of vectors.

4.1.1 Applications of Word Similarity

The concept of word similarity was traditionally captured within thesauri. A thesaurus is a lexicographic resource that specifies semantic relationships between words, listing for each word related words such as synonyms, hyponyms and hypernyms. Thesauri have been used to assist writers in selecting appropriate words and terms and in enriching the vocabulary of a text. To this end, modern word processors provide a thesaurus as a built in tool.

The area of IR has provided a new application for word similarity in the framework of query expansion. Good free-text retrieval queries are difficult to formulate since the same concept may be denoted in the text by different words and terms. Query expansion is a technique in which a query is expanded with terms that are related to the original terms that were given by the user, in order to improve the quality of the query. Various query expansion methods have been implemented, both by researchers and in commercial systems, that rely on manually crafted thesauri or on statistical measures for word similarity.

Word similarity may also be useful for disambiguation and language modeling in the area of NLP and speech processing. Many disambiguation methods and language models rely on word co-occurrence statistics that are used to estimate the likelihood of alternative interpretations of a natural language utterance (in speech or text). Due to data sparseness, though, the likelihood of many word co-occurrences cannot be estimated reliably from a corpus, in which case statistics about similar words may be helpful.

Consider for example the following utterances, which may be misinterpreted by a speech recognizer.

- a. The bear ran away.
- b. The pear ran away.

A typical language model may prefer the first utterance if the word co-occurrence bear ran was encountered in a training corpus while the alternative co-occurrence pear ran was not. However, due to data sparseness it is quite likely that neither of the two alternative interpretations was encountered in the training corpus. In such cases information about word similarity may be helpful. Knowing that bear is similar to other animals may help us collect statistics to support the hypothesis that animal names can precede the verb ran. On the other hand, the names of other fruits, which are known to be similar to the word pear, are not likely to precede this verb in any training corpus. This type of reasoning was attempted in various disambiguation methods, where the source of word

similarity was either statistical (Dagan *et al.*, 1993; Essen & Steinbiss, 1992; Grishman & Sterling, 1993; Grishman *et al.*, 1986; Karov & Edelman, 1996; Lin, 1997; Schütze, 1992, 1993) or a manually crafted thesaurus (Jiang & Conrath, 1997; Resnik, 1995).

It should be noted that while all the applications mentioned above are based on some notion of "word similarity" the appropriate type of similarity relationship might vary. A thesaurus intended for writing assistance should identify words that resemble each other in their meaning, like aircraft and airplane, which may be substituted for each other.

For query expansion, on the other hand, it is also useful to identify contextually related words, like aircraft and airline, which may both appear in relevant target documents. Co-occurrence-based disambiguation methods would benefit from identifying words that have similar co-occurrence patterns. These might be words that resemble each other in their meaning, but may also have opposite meanings, like increase and decrease.

4.1.2 Co-occurrence relations

In the corpus-based framework a word is represented by data about its joint co-occurrence with other words in the corpus. Different types of co-occurrence relationships have been examined in the literature, for computing word similarity as well as for other applications. These relationships may be classified into two general types: grammatical relations, which refer to the co-occurrence of words within specified syntactic relations, and non-grammatical relations, which refer to the co-occurrence of words within a certain distance (window) in the text. As will be discussed below, the types of relations used in a particular word similarity system will affect the types of similarity that will be identified.

4.1.2.1 Non-grammatical relations

Non-grammatical co-occurrence relations refer to the joint occurrence of words within a certain distance (window) in the text. This broad definition captures several sub-types of co-occurrence relations such as n-grams, directional and non-directional co-occurrence within small windows, and co-occurrence within large windows or within a document.

An n-gram is a sequence of n words that appear consecutively in the text. N-gram models are used extensively in language modeling for automatic speech recognition systems, as well as in other recognition and disambiguation tasks. In an n-gram model the probability of an occurrence of a word in a sentence is approximated by its probability of occurrence within a short sequence of n words. Typically sequences of two or three words (bigrams or trigrams) are used, and their probabilities are estimated from a large corpus. These probabilities are combined to estimate the a priori probability of alternative acoustic interpretations of the utterance in order to select the most probable interpretation.

The information captured by n-grams is, to a large extent, only an indirect reflection of lexical,

syntactic and semantic relationships in the language. This is because the production of consecutive sequences of words is a result of more complex linguistic structures. However, n-grams have been shown to have practical advantages for several reasons: it is easy to formulate probabilistic models for them, they are very easy to extract from a corpus, and, above all, they have proved to provide useful probability estimations for alternative readings of the input.

Word similarity methods that are based on bigram relationships were tried for addressing the data sparseness problem in n-gram language modeling (Dagan *et al.*, 1994; Essen & Steinbiss, 1992). Word-similarities that are obtained by n-gram data may reflect a mixture of syntactic, semantic, and contextual similarities, as these are the types of relationships represented by n-grams. Such similarities are suitable for improving an n-gram language model, which, by itself, mixes these types of information.

A co-occurrence of words within a relatively large window in the text suggests that both words are related to the general topic discussed in the text. This hypothesis will usually hold for frequent co-occurrences, that is, for pairs of words that often co-occur in the same text. A special case for this type of relationship is co-occurrence within the entire document, which corresponds to a maximal window size.

Co-occurrence within large windows was used in the work of Gale *et al.* (1992) on word-sense disambiguation. In this work co-occurrence within a maximal distance of 50 words in each direction was considered. A window of this size captures context words that identify the topic of discourse. Word co-occurrence within a wide context was used also for language modeling in speech recognition, where the occurrence of a word affects the probability of other words in the larger context. In the context of computing word similarity, co-occurrence within a large window may yield topical similarities between words that tend to appear in similar contexts.

Co-occurrence of words within a small window captures a mixture of grammatical relations and topical co-occurrences. Typically, only co-occurrence of content words is considered since these words carry most semantic information. Smadja (1993) used co-occurrence within a small window as an approximation for identifying significant grammatical relations without using a parser. His proposal relies on an earlier observation that 98% of the occurrences of syntactic relations relate words that are separated by at most five words within a single sentence (Martin *et al.*, 1983). Smadja (1993) used this fact to extract lexical collocations, and applied the extracted data to language generation and information retrieval. Dagan *et al.* (1993) use this type of data as a practical approximation for extracting syntactic relationships. To improve the quality of the approximation, the direction of co-occurrence is considered, distinguishing between co-occurrences with words that appear to the left or to the right of the given word. The extracted data is used to compute word similarities, which capture both semantic similarities, as when using grammatical relations, but also some topical similarities, as when using co-occurrence within a larger context.

Another variant of co-occurrence within a small window appears in the work of Brown *et al.* (1991). They use a part-of-speech tagger to identify relations such as “the first verb to the right” or “the

first noun to the left", and then use these relations for word-sense disambiguation in MT. This type of relationship provides a better approximation for syntactically motivated relations while relying only on a part of speech tagger, which is a simpler resource compared to syntactic parsers.

4.1.3 Asymmetric Word Similarities

New trends have recently emerged with the study of asymmetric measures (Michelbacher *et al.*, 2007). The idea of an Asymmetric Association Measures (AAM) is inspired by the fact that within the human mind, association between two words or concepts is not always symmetric. For pairs like fruit and apple, one would agree that there is a strong mutual association between the two. When thinking of fruit, it is not very far-fetched to think of apple as well and vice versa. There are other pairs, however, that do not exhibit this kind of strong association in both directions. Think of the pair fruit and mango, for example. Mango is probably not the first thing that comes to one's mind when hearing the word fruit. On the other hand, mango is strongly associated with the concept of a fruit. An example from Michelbacher *et al.* (2007) reads: *"there is a tendency for a strong forward association from a specific term like adenocarcinoma to the more general term cancer, whereas the association from cancer to adenocarcinoma is weak"*. According to Michelbacher *et al.* (2007), this idea bears some resemblance to the prototype theory (Rosch, 1973), where objects are regarded as members of different categories. Some members of the same category are more central than others making them more prototypical of the category they belong to. For instance, *cancer* would be more central than *adenocarcinoma*. However, we deeply believe that the main background for the direction of association lies in the notion of specific and general terms. Indeed, it is clear that there exists a tendency for a strong forward association from a specific term to the more general term but the backwards association is weaker. Within this scope, several recent works have proposed the use of asymmetric similarity measures. We believe that this idea has the potential to bring about significant improvements in the acquisition of word semantic relations.

4.1.3.1 Asymmetric Association Measures

Pattern-based measures can embody asymmetry as they were initially defined to discover the hypernymy/hyponymy relation. But, Ohshima & Tanaka (2009) is certainly the approach that makes the most of asymmetric patterns. Indeed, instantiating and sending to a search engine a number of patterns filled only with one possible candidate may guarantee the extraction of hypernymy/hyponymy or meronymy/holonymy relations if asymmetric patterns exist. However, we know that pattern-based measures are sensitive to word polysemy and pattern ambiguity. Moreover, they are language-dependent techniques which are difficult to replicate for different languages.
In order to stay within the domain of language-independent and unsupervised methodologies a

number of asymmetric association measures have been proposed (Pecina & Schlesinger, 2006; Tan *et al.*, 2004) and applied to the problems of taxonomy construction (Cleuziou *et al.*, 2010; Sanderson & Croft, 1999), cognitive psycholinguistics (Michelbacher *et al.*, 2007) and general-specific word order induction (Dias *et al.*, 2008).

Sanderson & Croft (1999) is certainly one of the first studies to propose the use of the conditional probability, Equation 4.1, for taxonomy construction.

$$P(x|y) = \frac{P(x,y)}{P(y)}. \tag{4.1}$$

They assume that a term t_2 subsumes a term t_1 if the documents in which t_1 occurs are a subset of the documents in which t_2 occurs constrained by $P(t_2|t_1) \geq 0.8$ and $P(t_1|t_2) < 1$. By gathering all subsumption relations, they build the semantic structure of any domain, which corresponds to a directed acyclic graph. In Sanderson & Lawrie (2000), the subsumption relation is relieved to the following expression $P(t_2|t_1) \geq P(t_1|t_2)$ and $P(t_2|t_1) > t$ where t is a given threshold and all term pairs found to have a subsumption relationship are passed through a transitivity module, which removes extraneous subsumption relationships in the way that transitivity is preferred over direct pathways, thus leading to a non-triangular directed acyclic graph.

Michelbacher *et al.* (2007) propose two different measures to model the notion of asymmetric association. Their intent is to determine to what extent these two measures of directed association can be used as a model for directed psychological association in the human mind. These two measures are the plain conditional probability and the ranking measure $R(.\|.)$ based on the Pearson's χ^2 test. In particular, let $t_i, i = 1 \ldots n$ be the list of all terms which co-occur with term t ordered with respect to the value $\chi^2(t, t_i)$. Then $R(t_i\|t)$ is the rank of term t_i in this list. The results were evaluated against a large number of free association norms, collected from human subjects, and they found that the measures were able to distinguish between highly symmetric and highly asymmetric pairs to some extent, but the overall accuracy in predicting the degree of asymmetry was low.

In the specific domain of word order discovery, Dias *et al.* (2008) propose a methodology based on directed graphs and the TextRank algorithm (Mihalcea & Tarau, 2004) to automatically induce a general-specific word order for a given vocabulary based on Web corpora frequency counts. A directed graph is obtained by keeping the edge, which corresponds to the maximum value of the asymmetric association measure between two words. Then, the TextRank is applied and produces an ordered list of nouns, on a continuous scale, from the most general to the most specific. Eight of the AAM used in that work will be evaluated in the context of asymmetric similarity between sentences: the Added Value (Equation 4.2), the Braun-Blanket (Equation 4.3), the Certainty Factor (Equation 4.4), the Conviction (Equation 4.5), the Gini Index (Equation 4.6), the J-measure (Equation 4.7), the Laplace (Equation 4.8) and the Conditional Probability (Equation 4.1).

$$AV(x\|y) = P(x|y) - P(x). \tag{4.2}$$

$$BB(x\|y) = \frac{f(x,y)}{f(x,y) + f(\bar{x},y)}. \tag{4.3}$$

$$CF(x\|y) = \frac{P(x|y) - P(x)}{1 - P(x)}. \tag{4.4}$$

$$CO(x\|y) = \frac{P(x) \times P(\bar{y})}{P(x,\bar{y})}. \tag{4.5}$$

$$\begin{aligned} GI(x\|y) = P(y) \times (P(x|y)^2 + P(\bar{x}|y)^2) - P(x)^2 \\ P(\bar{y}) \times (P(x|\bar{y})^2 + P(\bar{x}|\bar{y})^2) - P(\bar{x})^2. \end{aligned} \tag{4.6}$$

$$JM(x\|y) = P(x,y) \times \log \frac{P(x|y)}{P(x)} + P(\bar{x},y) \times \log \frac{P(\bar{x}|y)}{P(\bar{x})}. \tag{4.7}$$

$$LP(x\|y) = \frac{N \times P(x,y) + 1}{N \times P(y) + 2}. \tag{4.8}$$

4.1.3.2 Asymmetric Attributional Word Similarities

In Dias (2010) it was noted that it is unjustified from linguistic point of view to assume all the dimensions of a vector space model to be orthogonal to each other. Since each dimension typically corresponds to a context word, this is equivalent to the assumption that every two words denote disparate meanings. Apparently, such a vector space model fails to account adequately for contexts that are similar in meaning or synonymous.

The InfoSimba Similarity (IS) aims to measure the correlations between all the pairs of words in two word context vectors instead of just relying on their exact match as with the cosine similarity measure (Equation 4.9). Further, IS guarantees to catch similarity between pairs of words, even when they do not share contexts, due to data sparseness for example, nevertheless they have similar

contexts. It is defined in Equation 4.10 where $S(.,.)$ is any symmetric similarity measure and each W_{ij} corresponds to the attribute word at the jth position in the vector X_i, p is the length of the vector X_i.

$$cos(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik} \times X_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2} \times \sqrt{\sum_{k=1}^{p} X_{jk}^2}}. \tag{4.9}$$

$$IS(X_i, X_j) = \frac{\sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik} \times X_{il} \times S(W_{ik}, W_{il}) + \\ \sum_{k=1}^{p} \sum_{l=1}^{p} X_{jk} \times X_{jk} \times S(W_{jk}, W_{jl}) - \\ \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl}) \end{array} \right)}. \tag{4.10}$$

In the context of asymmetric attributional word similarities research (Freitag *et al.*, 2005; Lund *et al.*, 1995) the directions of co-occurrences is noted and exploited, but there does not exist an in-depth study neither a theoretical account of this phenomenon. The efforts are directed towards developing asymmetric distributional similarity measures such as the Kullback-Leibler divergence (Kullback & Leibler, 1951) defined in Equation 4.11 where $A = \{\langle z,r \rangle | \exists (x,z,r) \wedge \langle z,r \rangle | \exists (y,z,r)\}$, which has been regularly set apart from the Jensen-Shannon divergence (Menéndez *et al.*, 1997), its symmetric counterpart. We can also point at the cross entropy described in Pecina & Schlesinger (2006).

$$KL(x \| y) = \sum_{\langle z,r \rangle \in A} \log P(z|x) \times \frac{\log P(z|x)}{\log P(z|y)}. \tag{4.11}$$

Although there are many asymmetric similarity measures, they evidence problems that may reduce their utility. On the one hand, asymmetric association measures can only evaluate the generality/specificity relation between words that are known to be in a semantic relation such as in Sanderson & Croft (1999) and Dias *et al.* (2008). Indeed, they generally capture the direction of association between two words based on document contexts and only take into account a loose semantic proximity between words. For example, it is highly probable to find that *Apple* is more general than *iPad*, which can not be assimilated to an hypernymy/hyponymy or meronymy/holonymy relation. On the other hand, asymmetric attributional word similarities only take into account common contexts to assess the degree of asymmetric relatedness between two words. To leverage these issues, in AIS measure, which underlying idea is to say that one word x is semantically related to word y and x is more general than y, if x and y share as many relevant related words as possible and each context word of x is likely to be more general than most of the context words of y. The AIS is defined in Equation 4.12, where $AS(.\|.)$ is any asymmetric similarity measure, likewise for the IS in Equation

4.10 where $S(.,.)$ stands for any symmetric similarity measure. We also define its simplified version $AISs(.\|.)$ in 4.13.

$$AIS(X_i\|X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times AS(W_{ik}\|W_{il}) + \\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk} \times X_{jk} \times AS(W_{jk}\|W_{jl}) - \\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}) \end{array}\right)}. \tag{4.12}$$

$$AISs(X_i\|X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}) \leftrightarrow W_{ik} \neq W_{jl}. \tag{4.13}$$

Bellow we give a sample calculation of the simplified AIS, Equation 4.13, with *Added Value* measure, for the following pair of sentences:

- X_i : Rumen studies in Paris.
- X_j : Rumen studies in France.

Table 4.1: Web frequencies for calculations with ***All Words***

W_i	**Frequency**	W_j	**Frequency**
Rumen	14700000	*Rumen*	14700000
studies	261000000	*studies*	261000000
in	505400000	*in*	505400000
Paris	437000000	*France*	838000000

Table 4.2: Web frequencies for calculations with ***All Words***

		Frequency	
W_i	W_j	$W_i \cap W_j$	$W_j \cap W_i$
Rumen	*studies*	2080000	2080000
Rumen	*in*	15700000	15600000
Rumen	*France*	994000	994000
studies	*in*	10900000	10900000
studies	*France*	190000000	190000000
in	*France*	226000000	226000000
Paris	*Rumen*	688000	688000

Table 4.2: (continued)

		Frequency	
W_i	W_j	$W_i \cap W_j$	$W_j \cap W_i$
Paris	*studies*	154000000	166000000
Paris	*in*	132000000	132000000
Paris	*France*	318000000	315000000

$$
\begin{aligned}
&AISs(X_i \| X_j) = \\
&= \frac{1}{4 \times 4} \left(\begin{array}{l} \big((1 \times 1 \times AV(Rumen \| studies)) + (1 \times 1 \times AV(Rumen \| in)) + (1 \times 1 \times AV(Rumen \| France))\big) \times \\ \big((1 \times 1 \times AV(studies \| Rumen)) + (1 \times 1 \times AV(studies \| in)) + (1 \times 1 \times AV(studies \| France))\big) \times \\ \big((1 \times 1 \times AV(in \| Rumen)) + (1 \times 1 \times AV(in \| studies)) + (1 \times 1 \times AV(in \| France))\big) \times \\ \big((1 \times 1 \times AV(Paris \| Rumen)) + (1 \times 1 \times AV(Paris \| studies)) + \\ \quad (1 \times 1 \times AV(Paris \| in)) + (1 \times 1 \times AV(Paris \| France))\big) \end{array} \right) \\
&= \frac{1}{4 \times 4} \left(\begin{array}{l} \big(0.001 + 0.024 + (-0.006)\big) \times \big(0.020 + (-0.100) + 0.105\big) \times \\ \big(0.833 + (-0.194) + 0.034\big) \times \big(- 0.157 + 0.387 + 0.058 + 0.176\big) \end{array} \right) \\
&= 0.063 \times (0.019 \times 0.024 \times 0.673 \times 0.464) \\
&= 0.063 \times (0.000142) \\
&= \mathbf{0.0000089}
\end{aligned} \tag{4.14}
$$

$$
\begin{aligned}
&AISs(X_j \| X_i) = \\
&= \frac{1}{4 \times 4} \left(\begin{array}{l} \big((1 \times 1 \times AV(Rumen \| studies)) + (1 \times 1 \times AV(Rumen \| in)) + (1 \times 1 \times AV(Rumen \| Paris))\big) \times \\ \big((1 \times 1 \times AV(studies \| Rumen)) + (1 \times 1 \times AV(studies \| in)) + (1 \times 1 \times AV(studies \| Paris))\big) \times \\ \big((1 \times 1 \times AV(in \| Rumen)) + (1 \times 1 \times AV(in \| studies)) + (1 \times 1 \times AV(in \| Paris))\big) \times \\ \big((1 \times 1 \times AV(France \| Rumen)) + (1 \times 1 \times AV(France \| studies)) + \\ \quad (1 \times 1 \times AV(France \| in)) + (1 \times 1 \times AV(France \| Paris))\big) \end{array} \right) \\
&= \frac{1}{4 \times 4} \left(\begin{array}{l} \big(0.001 + 0.024 + (-0.005)\big) \times \big(0.020 + (-0.100) + 0.258\big) \times \\ \big(0.833 + (-0.194) + 0.067\big) \times \big(- 0.323 + 0.338 + 0.057 + 0.331\big) \end{array} \right) \\
&= 0.063 \times (0.020 \times 0.178 \times 0.706 \times 0.403) \\
&= 0.063 \times (0.000997) \\
&= \mathbf{0.0000628}
\end{aligned} \tag{4.15}
$$

4.2 Asymmetry between Words

Most of the metrics, which evaluate the degree of similarity between words are symmetric (Pecina & Schlesinger, 2006; Tan *et al.*, 2004), except perhaps pattern-based similarities Caraballo (1999);

Hearst (1992). Patterns can be helpful to learn knowledge from texts that can possibly be expressed by constructions known in advance and surely embody the easiest way to induce this knowledge. Most of the works in this area have been dealing with the identification of the hypernymy/hyponymy relation although some other word semantic relations such as synonymy and meronymy/holonymy have been tackled.

In order to extract hypernymy/hyponymy relations, Hearst (1992) first identifies a set of lexical-syntactic patterns that are easily recognizable (i.e. occur frequently and across text genre boundaries). These can be called seed patterns. Based on these seeds, she proposes a bootstrapping algorithm to semi-automatically acquire new more specific patterns such as *such NP as (NP,)* {or | and} NP*. Similarly, Caraballo (1999) uses predefined patterns such as *X is a (kind of) Y* or *X, Y, and other Zs*, following the discussion in Riloff & Shepherd (1997) that nouns in conjunctions or appositive relations tend to be semantically related. Despite the variety of approaches, two common characteristics are transversal to the methodology: (1) the necessity of manual effort as to compose the patterns and (2) the language-dependency of the method. Other drawbacks can be identified. In particular, lexical-syntactic patterns tend to be quite ambiguous as to which relations they indicate and this worsens when ambiguous words are involved. Also, mainly subsets of possible instances of semantic relations are likely to appear, thus imposing the existence of a great number of seed patterns. To overcome such drawbacks, new trends have recently emerged with the study of asymmetric measures Michelbacher *et al.* (2007).

The idea of an asymmetric measure is inspired by the fact that within the human mind, the association between two words or concepts is not always symmetric. For example, as stated in Michelbacher *et al.* (2007), "*there is a tendency for a strong forward association from a specific term like adenocarcinoma to the more general term cancer, whereas the association from cancer to adenocarcinoma is weak*". For instance, *cancer* would be more central than *adenocarcinoma*. Within this scope, seldom new researches have been emerging over the past few years, which propose the use of asymmetric similarity measures, which we believe can lead to great improvements in the acquisition of word semantic relations as shown in Cleuziou *et al.* (2011).

We present the eight asymmetric association measures used in this work that will be evaluated in the context of asymmetry between sentences: the Added Value (Equation 4.2), the Braun-Blanket (Equation 4.3), the Certainty Factor (Equation 4.4), the Conviction (Equation 4.5), the Gini Index (Equation 4.6), the J-measure (Equation 4.7), the Laplace (Equation 4.8), and the Conditional Probability (Equation 4.1).

4.3 Asymmetry between Sentences

There are a number of ways to compute the similarity between two sentences. Most similarity measures determine the distance between two vectors associated to two sentences (i.e. the vector space

model). However, when applying the classical similarity measures between two sentences, only the identical indexes of the row vector X_i and X_j are taken into account, which may lead to miscalculated similarities. To deal with this problem, different methodologies have been proposed, but the most promising one is certainly the one proposed by Dias *et al.* (2007), the InfoSimba informative similarity measure, expressed in Equation 4.10.

Although there are many asymmetric similarity measures between words, there does not exist any attributional similarity measure capable to assess whether a sentence is more specific/general than another one. To overcome this issue, we introduce the asymmetric InfoSimba similarity measure (AIS), which underlying idea is to say that a sentence T is semantically related to sentence H and H is more general than T (i.e. $T \rightarrow H$), if H and T share as many relevant related words as possible between contexts and each context word of H is likely to be more general than most of the context words of T. The AIS is defined in Equation 4.12.

As computation of the AIS may be hard due to orders of complexity, we also define its simplified version $AISs(.\|.)$ in Equation 4.13, which we will specifically use in our experiments.

As a consequence, an entailment ($T \rightarrow H$) will hold if and only if $AISs(T\|H) < AISs(H\|T)$. Otherwise, the entailment will not hold. This way, contrarily to existing methodologies, we do not need to define or tune thresholds. Indeed, due to its asymmetric definition, the asymmetric InfoSimba similarity measure allows to compare both sides of entailments.

4.4 Three Levels of Pre-Processing

In our work, we experienced three approaches for selecting the words for the calculation of the asymmetry between sentences.Thus we can assess which approach best performance to identify entailment by generality.In a first approach, we chose to do the calculations without restrictions, ie, do the calculations with all the words (see Figure 4.1).

The next approach was to use a list of stop words (for English[1] and for Portuguese[2]). Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. In computer search engines, a stop word is a commonly used word (such as “the”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. When building the index, most engines are programmed to remove certain words from any index entry. The list of words that are not to be added is called a stop word list (see table A.1). Stop words are deemed irrelevant for searching purposes because they occur frequently in the language for which the indexing engine has been tuned. In order to save both space and time, these words are dropped at indexing time and then ignored at search time. In the

[1]Source: `http://www.microsoft.com/en-us/download/confirmation.aspx?id=10024` [Last access: 14th December, 2013]

[2]Source: `http://www.linguateca.pt/chave/stopwords/` [Last access: 14th December, 2013]

context of our work, in this approach, the calculations are made with restrictions, ie, the words that are on the list of stop words are ignored (see Table A.1), are not considered in the calculations.(see Figure 4.2).

Finally, in this last approach, we introduce the concept of MWU (the next chapter explains in this concept), identified the MWU in sentences (see Table B.1) for the calculation of the asymmetry (see Figure 4.3).

In summary, our experiments are based on three approaches to the calculations, all words, using list of the Stop Words and finally using MWU. In the Chapter 5, we evaluate the performances of these approaches to identify which one best identifies entailment by generality.

4.4.1 Multiword Units Identification

Syntactical, statistical, hybrid syntactic-statistical, semantic and machine learning methodologies have been proposed to extract MWU. Although, there exists an important number of approaches, the identification of MWU still remains an open problem within an active research field. Historically, both syntactical and statistical approaches have been privileged. Purely linguistic systems follow the first part of the definition of MWU proposed in Choueka *et al.* (1983): a MWU *is defined as a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be directly derived from the meaning or connotation of its components*.

By definition, MWU are words that co-occur together more often than they would by chance in a given domain and usually convey conceptual information (Dias, 2002). For example, *tomber dans les pommes (to faint)* is a sequence of words which meaning is non-compositional i.e. it can not be reproduced by the sum of the meanings of its constituents and thus represents a typical MWU. MWU include a large range of linguistic phenomena as stated in Gross (1996), such as compound nouns (e.g. *chantier naval meaning* in French *shipyard*), phrasal verbs (e.g. *entrar em vigor meaning* in Portuguese *to come into force*), adverbial locutions (e.g. *sans cesse meaning* in French *constantly*), compound determinants (e.g. *un tas de meaning* in French *an amount of*), prepositional locutions (e.g. *au lieu de meaning* in French *instead of*), adjectival locutions (e.g. *a longo prazo meaning* in Portuguese *long-term*) and institutionalized phrases (e.g. *con carne*).

In our work, for extraction MWU in first five RTE dataset test (show list in Appendices B),used the Software for the Extraction of N-ary Textual Associations (SENTA) (Dias *et al.*, 1999), which is parameter free and language independent thus allowing the extraction of MWU from any raw text. It is based on the Mutual Expectation (ME) measure defined in Equation 4.16 and the GenLocalMaxs selection algorithm (is defined in Algorithm 1), which does not depend on any threshold. SENTA shows many advantages compared to different methodologies presented so far. It is parameter free, thus avoiding threshold tuning. It can extract relevant sequences of characters, thus allowing its application to character-based languages. And, interestingly, it obtains successful results for small texts as

it extracts MWU with low frequency with great accuracy without using lists of stop words or stemming.

$$ME(\widehat{S}) = \frac{n \times P(\widehat{S})^2}{\sum_{i_1=1}^{2} \cdots \sum_{i_{(n-1)}=i_{(n-2)}+1}^{n} P(p_{i_1 i_1} w_{i_1} p_{i_1 i_2} w_{i_2} \ldots p_{i_1 i_{(n-1)}} w_{i_{(n-1)}})}. \tag{4.16}$$

Algorithm 1 The GenLocalMaxs algorithm.

```
∀W_{n-1} ∈ Ω_{n-1}, ∀W_{n+1} ∈ Ω_{n+1}
if size(W) = 2 ∧ assoc(W) > assoc(W_{n+1}) then
    return MWU
else
    if size(W) ≠ 2 ∧ assoc(W) ≥ assoc(W_{n-1}) ∧ assoc(W) > assoc(W_{n+1}) then
        return MWU
    else
        return NO-MWU
    end if
end if
```

4.5 Sample of Calculation for Identify Entailment by Generality

In this chapter we present our methodology to identify entailment by generality between two sentences, we now apply our methodology on a pair of $T - H$ extracted from RTE-3 test set:

"<pair id="217" entailment="YES" task="IR" length="short" >
<t>Pierre Beregovoy, apparently left no message when he shot himself with a borrowed gun.</t>
<h>Pierre Beregovoy commits suicide.</h>
</pair>"

The AAM, we use in this demonstration is the Conditional Probability (Equation 4.1), for the calculations of the three approaches,the terms and their web frequencies are in the Appendix D. For the calculations used the *Google API*[1] to calculate all joint and marginal frequencies, so, instead of relying on a closed corpus and exact frequencies, we based our analysis on the Web and Web hits i. e. estimated number of documents where words appear - each pair needs approximately 17 minutes to get all the frequencies used in the respective AAM. A total of 5790669 queries derived from pairs of the first five RTE Challenges, which are submitted to *Google API* to know its frequency.
The following figures show the links between the terms in the three approaches.

[1] `https://code.google.com/apis/console/` [Last access: 14th December, 2013]

The next figure illustrates the links between sentence, when the calculations are made with all terms - *All Words* that compose theses sentences.

Figure 4.1: Sample calc with all words.

In this case, $AISs(T\|H) = 0.26$ and $AISs(H\|T) = 0.20$, then $AISs(T\|H) = 0.26 > AISs(H\|T) = 0.20$, so we conclude that T no entail H.

In the approach we use a list of stop words, in this example,for calculations exclude the T the words: *"no"*, *"he"*, *"with"* and *"a"*. H not exclude words because none of them are on the list of stop words (see Table A.1).

Figure 4.2: Sample calc with list of the stop words .

In this case, $AISs(T\|H) = 0.17$ and $AISs(H\|T) = 0.24$, then $AISs(T\|H) = 0.17 < AISs(H\|T) = 0.24$, so we conclude that T entail H - $(T- > H)$.

Through the concept of MWU, gives us the possibility of the links are not only in single words but also in terms, as we see below (Figure 4.3), the example has two MWU -*"Pierre Beregovoy"* and *"with a"*.

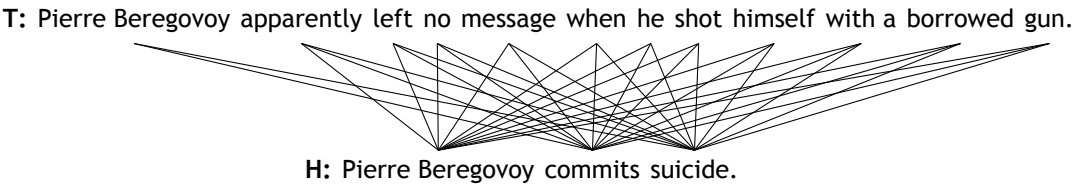

Figure 4.3: Sample calc with Multiword Units.

In this case, $AISs(T\|H) = 0.17$ and $AISs(H\|T) = 0.22$, then $AISs(T\|H) = 0.17 < AISs(H\|T) = 0.22$, so we conclude that T entail H - $(T- > H)$.

Chapter 5

Evaluating the Performance of our Methodology

"You may never know what results come of your action,
but if you do nothing there will be no result."
Mahatma Gandhi

In this Chapter we present the results obtained through the calculations of measures for evaluating the performance of our methodology. Analyze and compare the precisions and accuracies calculated.

5.1 Evaluation Scheme

With the evaluation the performance of our methodology will help us define what approach - with all words; without stop words; with MWU - and what AAM - the Added Value (Equation 4.2), the Braun-Blanket (Equation 4.3), the Certainty Factor (Equation 4.4), the Conviction (Equation 4.5), the Gini Index (Equation 4.6), the J-measure (Equation 4.7), the Laplace (Equation 4.8), and the Conditional Probability (Equation 4.1), what better way to recognize entailment by generality.
With this new definition, we know how to implement future framework and / or toolkits unsupervised and language-independent, with different objectives in NLP.
Our evaluation is based on analysis of the results obtained through the measures that we present below. The calculation of these measures are based on the Confusion Matrix (CM).

5.1.1 Measures to evaluate the performance

Classification or categorization is the task of assigning objects from a universe to two or more classes or categories. In the field of AI, a CM is a visualization tool typically used in supervised and unsupervised learning. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a CM is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

When the dataset is unbalanced (when the number of samples in different classes vary greatly) the error rate of a classifier is not representative of the true performance of the classifier.
The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct **predictions** that an instance is **Entailment**,
- b is the number of incorrect **predictions** that an instance is **No Entailment**,
- c is the number of incorrect **predictions** that an instance in **Entailment** and
- d is the number of correct **predictions** that an instance is **No Entailment**.

	YES is correct	**NO is correct**
YES was assigned	**a**	**b**
NO was assigned	**c**	**d**

Table 5.1: Contingency table for evaluating a binary classifier. For example, a is the number of objects in the category of interest that were correctly assigned to the category. (Manning & Schütze, 1999)

For binary classification, classifiers are typically evaluated using a table of counts like table 5.1. An important measure is classification Accuracy (AC) and Precision (P) which is defined in equation 5.1 and equation 5.2, respectively.
The AC is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d} \tag{5.1}$$

P is defined as a measure of the proportion of selected items that the system got right:

$$P = \frac{a}{a+b} \tag{5.2}$$

To evaluate the overall performance of our experiments, we used two types of averaging of the previous measures, calculate the arithmetic average and the calculation of the weighted average[1].
The weighted average formula is used to calculate the average value of a particular set of numbers with different levels of relevance. The relevance of each number is called its weight. The weights should be represented as a percentage of the total relevancy. Therefore, all weights should be equal to 100%, or 1. The most common formula used to determine an average is the arithmetic mean formula. This formula adds all of the numbers and divides by the amount of numbers. For example the average of 1, 2 and 3 would be the sum $1+2+3$ divided by 3, which would return 2. However,

[1]The weighted average is a average where there is some variation in the relative contribution of individual data values to the average. Each data value (X_i) has a weight assigned to it (W_i). Data values with larger weights contribute more to the weighted average and data values with smaller weights contribute less to the weighted average.

the weighted average formula looks at how relevant each number is. Say that 1 only happens 10% of the time while 2 and 3 each happen 45% of the time. The percentages in this example would be the weights. The weighted average would be $2,35$.
More specifically, in our work, we defined following equations - *Average Accuracy* (Equation 5.3), *Average Precision* (Equation 5.4) and *Weighted Average Accuracy* (Equation 5.5), *Weighted Average Precision* (Equation 5.6).

$$\overline{AC} = \frac{\sum_{i=1}^{n} AC_i}{n} \tag{5.3}$$

$$\overline{P} = \frac{\sum_{i=1}^{n} P_i}{n} \tag{5.4}$$

$$\overline{AC}_w = \frac{\sum_{i=1}^{n} AC_i W_i}{\sum_{i=1}^{n} W_i} \tag{5.5}$$

$$\overline{P}_w = \frac{\sum_{i=1}^{n} P_i W_i}{\sum_{i=1}^{n} W_i} \tag{5.6}$$

5.2 All pairs of the Test Set of the first five RTE Challenges

In this section, we present the performance of our methodology measured in terms **Arithmetic Average** and **Weighted Average** of both **Accuracy** and **Precision**, as these two metrics are commonly used to assess performance in the RTE field. With the results presented here we can see how our methodology behaves in the RTE task. We detail our results over five sets of $T - H$ pairs used for evaluation in the early RTE challenges. This endeavor provides for a fair comparison of our methodology and the results of other researchers.

5.2.1 All Words

This approach involves more calculations, compared to the other two approaches, since all the words in the texts snippets are considered, as we demonstrate in Section 4.5. More words imply also more search requests to the *Google* API, consequently longer delay in obtaining results.
In terms of **Accuracy** the measure that achieves the best results is the *Braun-Blanket* with **Arithmetic Average** of $0,55$ and **Weighted Average** equal to $0,54$. The *J-measure* in **Weighted Average** achieves $0,54$. The worst performance is of the *Conviction* measure with $0,52$ for both **Arithmetic Average**

and **Weighted Average**. Similarly, the *J-measure* achieves $0,52$ in **Weighted Average**.

Table 5.2: Accuracy Average by RTE Challenges | With All Words

ACCURACY by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,53	0,52	0,51	0,58	0,55	0,54
BRAUN-BLANKET	0,53	0,51	0,52	0,59	0,58	0,55
CERTAINTY FACTOR	0,51	0,52	0,52	0,56	0,56	0,53
CONDITIONAL PROBABILITY	0,50	0,52	0,51	0,58	0,55	0,53
CONVICTION	0,48	0,50	0,49	0,57	0,53	0,52
GINI INDEX	0,48	0,52	0,53	0,58	0,57	0,54
J-MEASURE	0,52	0,50	0,51	0,60	0,53	0,53
LAPLACE	0,50	0,52	0,51	0,56	0,55	0,53
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,52	-	0,51	0,58	-	0,53
BRAUN-BLANKET	0,51	-	0,52	0,58	-	0,54
CERTAINTY FACTOR	0,51	-	0,52	0,55	-	0,53
CONDITIONAL PROBABILITY	0,50	-	0,51	0,57	-	0,53
CONVICTION	0,49	-	0,49	0,57	-	0,52
GINI INDEX	0,48	-	0,53	0,59	-	0,53
J-MEASURE	0,51	-	0,51	0,61	-	0,54
LAPLACE	0,50	-	0,51	0,56	-	0,52

Table 5.2 reveals that the results fit a short range between $0,52$ and $0,55$ for the **Arithmetic Average** and between $0,52$ and $0,54$ for the **Weighted Average**. Analysis with respect to different challenges shows that globally **AAM**s perform best on **RTE-4** data while **RTE-1** data seems to be the most challenging set.

In spite of the low values of **Accuracy**, in **Precision - Entailment** for this approach the *J-measure* stands out with $0,81$ and $0,73$ for **Arithmetic Average** and **Weighted Average**, respectively. With respect to the *J-measure* in **Arithmetic Average**, we highlight the excellent result achieved on the **RTE-2** data set, namely $0,91$. In **Precision - Entailment** the worst measure is *Braun-Blanket* with $0,43$ and $0,44$ for **Arithmetic Average** and **Weighted Average**, respectively.

The averages that we present in Table 5.3 do not have the same behavior as the averages we presented in Table 5.2. The best result in terms of **Precision - Entailment** is much higher compared to the second best result. With all words the best result is achieved by the *J-measure* for **RTE-2** set.

Its worst performance is on the **RTE-1** test set.

Table 5.3: PRECISION - ENTAILMENT Average by RTE Challenges | With All Words

PRECISION - ENTAILMENT by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,48	0,85	0,63	0,69	0,67	0,66
BRAUN-BLANKET	0,43	0,62	0,47	0,57	0,57	0,53
CERTAINTY FACTOR	0,47	0,80	0,58	0,67	0,63	0,63
CONDITIONAL PROBABILITY	0,44	0,83	0,61	0,67	0,66	0,64
CONVICTION	0,53	0,58	0,60	0,63	0,64	0,60
GINI INDEX	0,53	0,81	0,63	0,70	0,69	0,67
J-MEASURE	0,52	0,91	0,84	0,87	0,89	0,81
LAPLACE	0,44	0,83	0,61	0,72	0,66	0,65
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,49	-	0,64	0,67	-	0,60
BRAUN-BLANKET	0,44	-	0,48	0,56	-	0,49
CERTAINTY FACTOR	0,48	-	0,60	0,65	-	0,58
CONDITIONAL PROBABILITY	0,45	-	0,63	0,65	-	0,58
CONVICTION	0,54	-	0,59	0,64	-	0,59
GINI INDEX	0,53	-	0,64	0,70	-	0,62
J-MEASURE	0,51	-	0,84	0,86	-	0,73
LAPLACE	0,45	-	0,63	0,71	-	0,59

The results for **Precision - No Entailment** show that the best measure is the *Braun-Blanket* with significant differences for the second best measure (contrary to what we have seen in table 5.3). With the bests results in **RTE-1** and the worse results in **RTE-2**, with emphasis on the *J-measure* that has obtained $0,09$.

Considering the results for **All Words**, Table 5.2, we can not conclude that our methodology is capable to recognize textual entailment as the best **Accuracy**, achieved by *Braun-Blanket*, is as low as $0,55$. On the other side, when we analyze Table 5.3 and Table 5.4 we can conclude that our methodology identifies better **Entailment** compared to recognition of **No Entailment** cases with *J-measure* and *Braun-Blanket*, respectively.

An interesting pattern to observe is a symmetric behavior with respect to **Precision**. Thus, the **RTE-2** is the challenge that presents the best results for **Precision - Entailment**, but this is the one that has worst results for **Precision - No Entailment**, and the **RTE-1** is the challenge that has the

worst results for **Precision - Entailment**, but the best results for **Precision - No Entailment**.

Table 5.4: PRECISION - NO ENTAILMENT Average by RTE Challenges | With All Words

PRECISION - NO ENTAILMENT by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,57	0,20	0,36	0,47	0,42	0,40
BRAUN-BLANKET	0,62	0,40	0,56	0,61	0,59	0,56
CERTAINTY FACTOR	0,55	0,24	0,43	0,44	0,48	0,43
CONDITIONAL PROBABILITY	0,57	0,22	0,37	0,50	0,43	0,42
CONVICTION	0,43	0,42	0,40	0,52	0,42	0,44
GINI INDEX	0,42	0,24	0,40	0,47	0,44	0,39
J-MEASURE	0,52	0,09	0,17	0,34	0,17	0,26
LAPLACE	0,56	0,22	0,37	0,40	0,43	0,40
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,63	-	0,37	0,49	-	0,49
BRAUN-BLANKET	0,67	-	0,57	0,61	-	0,62
CERTAINTY FACTOR	0,61	-	0,44	0,45	-	0,50
CONDITIONAL PROBABILITY	0,63	-	0,38	0,49	-	0,50
CONVICTION	0,50	-	0,39	0,49	-	0,46
GINI INDEX	0,49	-	0,41	0,48	-	0,46
J-MEASURE	0,58	-	0,17	0,36	-	0,37
LAPLACE	0,62	-	0,38	0,41	-	0,47

5.2.2 Without Stop Words

As already mentioned, in this approach we have excluded from calculations words that do not add relevant information (*stop words*) to the sentences that compose the pair. Compared to the previous approach, this one performs fewer calculations, due to the exclusion of about half of the terms.

In this approach, our methodology shows a different behavior. Regarding **Arithmetic Average** of **Accuracy** the best measures are *Braun-Blanket*, *Certainty Factor*, *Conditional Probability* and *Laplace* with $0,53$. For **Weighted Average** of **Accuracy** the best measures are the *Conditional Probability* and the *Certainty Factor*, both with $0,53$. The worst results are achieved by *Conviction*, *Gini Index* and *J-Measure* reaching as low as $0,49$.

As happened in the approach **With all Words**, also in this approach for **Accuracy**, there is no significant difference between the measurements, as we confirmed in the results shown in Table 5.5.

However, the **RTE-4** set allows for relatively better performance compared to **RTE-1** and **RTE-2** data sets.

Table 5.5: Accuracy Average by RTE Challenges | Without Stop Words

ACCURACY by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,50	0,52	0,52	0,55	0,54	0,52
BRAUN-BLANKET	0,52	0,50	0,52	0,54	0,56	0,53
CERTAINTY FACTOR	0,50	0,51	0,51	0,57	0,54	0,53
CONDITIONAL PROBABILITY	0,51	0,52	0,52	0,56	0,54	0,53
CONVICTION	0,51	0,53	0,49	0,50	0,54	0,51
GINI INDEX	0,51	0,50	0,50	0,55	0,52	0,51
J-MEASURE	0,50	0,49	0,51	0,53	0,52	0,51
LAPLACE	0,51	0,52	0,52	0,53	0,55	0,53
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,50	-	0,52	0,54	-	0,52
BRAUN-BLANKET	0,52	-	0,52	0,53	-	0,52
CERTAINTY FACTOR	0,50	-	0,51	0,57	-	0,53
CONDITIONAL PROBABILITY	0,51	-	0,52	0,55	-	0,53
CONVICTION	0,51	-	0,49	0,50	-	0,50
GINI INDEX	0,51	-	0,50	0,55	-	0,52
J-MEASURE	0,50	-	0,51	0,53	-	0,51
LAPLACE	0,51	-	0,52	0,52	-	0,52

With respect to **Precision - Entailment**, the measure with best performance is the *J-measure*. On the **RTE-2** data set it achieves $0,89$ as evidenced by Table 5.6. The *J-measure* achieves $0,75$ in **Arithmetic Average**, while the second best measure, *ADDED VALUE*, only achieves $0,65$. In **Weighted Average**, *J-measure* achieves $0,66$ and the second best measure, *ADDED VALUE*, achieves $0,59$. On the other hand, in average, the *Conviction* and *Gini Index* have the worst performances. Individually, the **AAM**'s have the best performance on **RTE-2** dataset and the worst results are obtained on **RTE-1** dataset.

Table 5.6: PRECISION - ENTAILMENT Average by RTE Challenges | Without Stop Words

PRECISION - ENTAILMENT by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,44	0,80	0,66	0,68	0,71	0,65
BRAUN-BLANKET	0,41	0,72	0,59	0,61	0,64	0,60
CERTAINTY FACTOR	0,44	0,74	0,60	0,66	0,65	0,62
CONDITIONAL PROBABILITY	0,43	0,78	0,65	0,68	0,68	0,64
CONVICTION	0,55	0,55	0,50	0,53	0,55	0,54
GINI INDEX	0,47	0,62	0,54	0,56	0,57	0,55
J-MEASURE	0,42	0,89	0,79	0,79	0,83	0,75
LAPLACE	0,43	0,78	0,65	0,65	0,70	0,64
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,45	-	0,66	0,66	-	0,59
BRAUN-BLANKET	0,42	-	0,60	0,60	-	0,54
CERTAINTY FACTOR	0,45	-	0,61	0,66	-	0,57
CONDITIONAL PROBABILITY	0,43	-	0,66	0,67	-	0,58
CONVICTION	0,54	-	0,50	0,53	-	0,53
GINI INDEX	0,48	-	0,54	0,56	-	0,52
J-MEASURE	0,42	-	0,80	0,78	-	0,66
LAPLACE	0,43	-	0,66	0,63	-	0,57

Unusual behavior is observed with respect to **Precision - No Entailment**, Table 5.7. Although the results are not significantly different, the measure that achieves the best value in **Arithmetic Average**, is not the same measures that achieve the best values in the **Weighted Average**. The **Arithmetic Average Precision - No Entailment** of the *Conviction* is $0,49$, while for **Weighted Average** we have two measure with best results, *Braun-Blanket* and *Gini Index* with $0,53$ of precision. The *J-measure* shows the worst performance in both averages. The best averages are achieved with *Braun-Blanket* ($0,62$ and $0,7$ respectively).

Table 5.7: PRECISION - NO ENTAILMENT Average by RTE Challenges | Without Stop Words

PRECISION - NO ENTAILMENT by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,56	0,24	0,36	0,43	0,36	0,39
BRAUN-BLANKET	0,62	0,28	0,44	0,47	0,48	0,46
CERTAINTY FACTOR	0,56	0,28	0,41	0,48	0,44	0,44
CONDITIONAL PROBABILITY	0,60	0,26	0,37	0,44	0,41	0,41
CONVICTION	0,46	0,51	0,47	0,46	0,54	0,49
GINI INDEX	0,54	0,38	0,46	0,54	0,47	0,48
J-MEASURE	0,57	0,09	0,21	0,28	0,21	0,27
LAPLACE	0,59	0,26	0,37	0,42	0,39	0,41
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,64	-	0,37	0,42	-	0,48
BRAUN-BLANKET	0,70	-	0,45	0,45	-	0,53
CERTAINTY FACTOR	0,63	-	0,41	0,48	-	0,51
CONDITIONAL PROBABILITY	0,67	-	0,37	0,43	-	0,49
CONVICTION	0,53	-	0,47	0,47	-	0,49
GINI INDEX	0,61	-	0,46	0,53	-	0,53
J-MEASURE	0,65	-	0,21	0,28	-	0,38
LAPLACE	0,67	-	0,37	0,41	-	0,48

The same pattern is observed here with respect to **Precision** as in the approach **With All Words**, namely the **RTE-2** is the challenge that presents the best results for **Precision - Entailment**, but this is the one that has worst results for **Precision - No Entailment**, and the **RTE-1** is the challenge that has the worst results for **Precision - Entailment**, but the best results for **Precision - No Entailment**. In this approach *Braun-Blanket* achieves the best performance.

In summary, the best **Accuracy** is achieved by three measures - *Braun-Blanket*, *Certainty Factor* and *Conditional Probability*, while the best **Precisions** values are achieved by *Braun-Blanket*, *Conviction*, *Gini Index* and *J-measure*.

5.2.3 With Multiword Units

Due to the use of MWU this approach requires fewer calculations than in the first approach. Compared to the previous approach, the number of calculations is roughly equivalent.

We expect that the use of MWU would allow for improved results. However, with regard to **Accuracy**, this is not the case as we have performance figures that are not significantly different from the previous approaches. The measure with best performance is the *Braun-Blanket*, with values of $0,54$ for **Arithmetic Average** and $0,56$ for the **Weighted Average**, also the *Laplace* obtained $0,56$ for the **Weighted Average**.

Table 5.8: Accuracy Average by RTE Challenges | With MWU

ACCURACY by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,50	0,51	0,52	0,57	0,57	0,53
BRAUN-BLANKET	0,53	0,51	0,53	0,55	0,58	0,54
CERTAINTY FACTOR	0,52	0,51	0,51	0,54	0,55	0,53
CONDITIONAL PROBABILITY	0,51	0,51	0,52	0,57	0,55	0,53
CONVICTION	0,48	0,48	0,51	0,55	0,54	0,51
GINI INDEX	0,49	0,51	0,52	0,58	0,56	0,53
J-MEASURE	0,51	0,49	0,51	0,56	0,53	0,52
LAPLACE	0,52	0,51	0,52	0,56	0,56	0,53
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,50	-	0,52	0,57	-	0,53
BRAUN-BLANKET	0,61	-	0,53	0,55	-	0,56
CERTAINTY FACTOR	0,51	-	0,51	0,53	-	0,52
CONDITIONAL PROBABILITY	0,51	-	0,52	0,57	-	0,53
CONVICTION	0,48	-	0,51	0,55	-	0,51
GINI INDEX	0,49	-	0,52	0,57	-	0,53
J-MEASURE	0,50	-	0,51	0,56	-	0,52
LAPLACE	0,61	-	0,52	0,56	-	0,56

In particular, these last two measures, *Braun-Blanket* and *Laplace*, achieve $0,61$ in **RTE-1** for **Weighted Average**. In this case, our methodology has better performance compared to the methodologies presented in Section 2.2, namely, Bayer *et al.* (2005), Glickman & Dagan (2005) and Perez *et al.* (2005), as they obtained $0,586$, $0,586$ and $0,495$. The **RTE-4** and **RTE-5** data sets afford for the best performance in this approach. **RTE-1** and **RTE-2** data sets, likewise in the previous approaches, exhibit low performance.

For **Precision - Entailment**, in this approach, the measure that stands out is the *Added Value*, with a significant difference compared to the second best measures (*Braun-Blanket* and *J-measure*). The

best result per challenge is obtained with the *Added Value* on **RTE-2**, achieving precision of $0,9$. The worst averages precision ($0,46$) is obtained with *Certainty Factor*.
The behavior of the measures was very similar in all RTE Challenges.

Table 5.9: PRECISION - ENTAILMENT Average by RTE Challenges | With MWU

PRECISION - ENTAILMENT by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,78	0,90	0,72	0,73	0,78	0,78
BRAUN-BLANKET	0,55	0,77	0,57	0,63	0,63	0,63
CERTAINTY FACTOR	0,49	0,45	0,43	0,46	0,46	0,46
CONDITIONAL PROBABILITY	0,43	0,44	0,46	0,55	0,49	0,48
CONVICTION	0,46	0,51	0,56	0,56	0,61	0,54
GINI INDEX	0,45	0,49	0,55	0,59	0,58	0,53
J-MEASURE	0,57	0,55	0,67	0,68	0,67	0,63
LAPLACE	0,44	0,44	0,46	0,52	0,49	0,47
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,79	-	0,73	0,71	-	0,74
BRAUN-BLANKET	0,55	-	0,58	0,61	-	0,58
CERTAINTY FACTOR	0,49	-	0,44	0,44	-	0,46
CONDITIONAL PROBABILITY	0,43	-	0,47	0,54	-	0,48
CONVICTION	0,48	-	0,55	0,57	-	0,53
GINI INDEX	0,45	-	0,55	0,58	-	0,53
J-MEASURE	0,56	-	0,67	0,68	-	0,63
LAPLACE	0,44	-	0,47	0,51	-	0,48

When we analyze Table 5.10, three measures stand out with good results, namely *Certainty Factor*, *Conditional Probability* and *Laplace*, where the latter two achieve a maximum precision of $0,67$. The worst results are obtained with *Added Value* ($0,21$ and $0,23$ on **RTE-1**) which, on the other hand, obtained strong results in **Precision - Entailment** (see Table 5.9). The **RTE-5** data set affords the best performance.

Table 5.10: PRECISION - NO ENTAILMENT Average by RTE Challenges | With MWU

PRECISION - NO ENTAILMENT by RTE Challenges						
AAM	**Arithmetic Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,21	0,11	0,29	0,42	0,36	0,28
BRAUN-BLANKET	0,50	0,26	0,47	0,47	0,53	0,45
CERTAINTY FACTOR	0,54	0,58	0,59	0,62	0,63	0,59
CONDITIONAL PROBABILITY	0,59	0,57	0,57	0,60	0,62	0,59
CONVICTION	0,49	0,45	0,47	0,53	0,48	0,48
GINI INDEX	0,52	0,53	0,49	0,56	0,54	0,53
J-MEASURE	0,45	0,44	0,35	0,45	0,39	0,42
LAPLACE	0,59	0,57	0,57	0,61	0,63	0,59
AAM	**Weighted Average**					
	RTE-1	**RTE-2**	**RTE-3**	**RTE-4**	**RTE-5**	**Average**
ADDED VALUE	0,23	-	0,31	0,43	-	0,32
BRAUN-BLANKET	0,56	-	0,48	0,48	-	0,51
CERTAINTY FACTOR	0,61	-	0,59	0,62	-	0,60
CONDITIONAL PROBABILITY	0,67	-	0,57	0,59	-	0,61
CONVICTION	0,55	-	0,47	0,52	-	0,52
GINI INDEX	0,61	-	0,49	0,56	-	0,55
J-MEASURE	0,51	-	0,35	0,44	-	0,43
LAPLACE	0,67	-	0,57	0,61	-	0,62

5.2.4 Summary

The purpose of this section is to evaluate our methodology against well known test data used to compare a number of methodologies. Although the obtained results are not excellent, they are promising and encouraging.

An individual analysis of Table 5.2 and Table 5.5, considering per challenge results, we conclude that **RTE-4** is associated to improved accuracy for approach **All Words** and **Without Stop Words** compared to other data sets. The best accuracy of the approach **With MWU** is achieved on **RTE-4** and **RTE-5** data sets, as evidenced in Table 5.8.

With respect to **Precision - Entailment**, Table 5.3, Table 5.6 and Table 5.9, **RTE-2** affords best results with **All Words** and **Without Stop Words** approaches, while **With Multiword Units** approach works best on **RTE-5** data set.

With respect to **Precision - No Entailment**, Table 5.4, Table 5.7 and Table 5.10, **RTE-1** affords best results with **All Words** and **Without Stop Words**. Similarly to **Precision - Entailment**, the best

Precision - No Entailment With Multiword Units approach is achieved on **RTE-5** data set.

Regarding the **Arithmetic Average** (Table 5.11), the combination that has the best performance is the *Braun-Blanket* measure on **All Words**. Best **Weighted Average** is achieved on **With WMU** approach by *Braun-Blanket* and *Laplace* measures. Overall, the worst result was obtained with the measure *Conviction* in approach **Without Stop Words**.

As already mentioned earlier, **Accuracy** values of our experiments on RTE Challenges span a relatively short range between $0,50$ and $0,56$.

Table 5.11: Accuracy Averages | Measures versus Approach

Averages ACCURACY by RTE Challenges | Measures versus Approach			
AAM	**Arithmetic Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0,54	0,52	0,53
BRAUN-BLANKET	0,55	0,53	0,54
CERTAINTY FACTOR	0,53	0,53	0,53
CONDITIONAL PROBABILITY	0,53	0,53	0,53
CONVICTION	0,52	0,51	0,51
GINI INDEX	0,54	0,51	0,53
J-MEASURE	0,53	0,51	0,52
LAPLACE	0,53	0,53	0,53
AAM	**Weighted Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0,53	0,52	0,53
BRAUN-BLANKET	0,54	0,52	0,56
CERTAINTY FACTOR	0,53	0,53	0,52
CONDITIONAL PROBABILITY	0,53	0,53	0,53
CONVICTION	0,52	0,50	0,51
GINI INDEX	0,53	0,52	0,53
J-MEASURE	0,54	0,51	0,52
LAPLACE	0,52	0,52	0,56

Table 5.11 points out the approach **Without Stop Words** as the one with worst performance in terms of accuracy, while **All Words** achieves slightly better accuracy compared to **With MWU**.

The combination with the best performance on the **Arithmetic Average Precision** is the measure *J-measure* with approach **All Words**. For the **Weighted Average Precision**, *Added Value* shows the best result **With MWU**. The worst result is obtained with measure *Conviction* **With MWU** - $0,46$.

Table 5.12: PRECISION - ENTAILMENT Averages | Measures versus Approach

Average PRECISION - ENTAILMENT by RTE Challenges | Measures versus Approach			
AAM	**Arithmetic Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0,66	0,65	0,78
BRAUN-BLANKET	0,53	0,60	0,63
CERTAINTY FACTOR	0,63	0,62	0,46
CONDITIONAL PROBABILITY	0,64	0,64	0,48
CONVICTION	0,60	0,54	0,54
GINI INDEX	0,67	0,55	0,53
J-MEASURE	0,81	0,75	0,63
LAPLACE	0,65	0,64	0,47
AAM	**Weighted Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0,60	0,59	0,74
BRAUN-BLANKET	0,49	0,54	0,58
CERTAINTY FACTOR	0,58	0,57	0,46
CONDITIONAL PROBABILITY	0,58	0,58	0,48
CONVICTION	0,59	0,53	0,53
GINI INDEX	0,62	0,52	0,53
J-MEASURE	0,73	0,66	0,63
LAPLACE	0,59	0,57	0,48

With respect to **Precision - Entailment** criterion, the approach that achieves the best results is **With All Words**.

In contrast to the results for **Precision - Entailment**, our method shows unsatisfactory behavior when considered from the perspective of **Precision - No Entailment**. For **Arithmetic Average** the best combination is *Certainty Factor*, *Conditional Probability* and *Laplace* **With MWU**. For **Weighted Average**, *Laplace* has the best performance **With MWU** approach. Note the low results obtained by *J-measure* and *Added Value*. In Table 5.13 the approach with the best performance is **With MWU**, and the worst performing approach is **Without Stop Words**.

After an exhaustive study and analysis of the results obtained from the application of our methodology, we can compare our results with the results of the methodologies presented in Section 2.2, namely, Bayer *et al.* (2005), Glickman & Dagan (2005) and Perez *et al.* (2005). They obtained $0,586$, $0,586$ and $0,495$ of accuracy, respectively. We prove that our methodology has better performance than was possible in previous works. On RTE-1 Challenge *With MWU* approach, our methodology achieved its best results. Table 5.8 shows that the measures *Braun-Blanket* and *Laplace* achieve

good results in **Weighted Average Accuracy**, namely $0,61$.

Table 5.13: PRECISION - NO ENTAILMENT Averages | Measures versus Approach

Average PRECISION - NO ENTAILMENT by RTE Challenges | Measures versus Approach			
AAM	**Arithmetic Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0,40	0,39	0,28
BRAUN-BLANKET	0,56	0,46	0,45
CERTAINTY FACTOR	0,43	0,44	0,59
CONDITIONAL PROBABILITY	0,42	0,41	0,59
CONVICTION	0,44	0,49	0,48
GINI INDEX	0,39	0,48	0,53
J-MEASURE	0,26	0,27	0,42
LAPLACE	0,40	0,41	0,59
AAM	**Weighted Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0,49	0,48	0,32
BRAUN-BLANKET	0,62	0,53	0,51
CERTAINTY FACTOR	0,50	0,51	0,60
CONDITIONAL PROBABILITY	0,50	0,49	0,61
CONVICTION	0,46	0,49	0,52
GINI INDEX	0,46	0,53	0,55
J-MEASURE	0,37	0,38	0,43
LAPLACE	0,47	0,48	0,62

5.3 Corpus TE by Generality

In this section, we present the results of an experiment designed to measure the degree to which our methodology is capable to distinguish TE by Generality. To this end we needed a corpus of entailment instances that were labeled either TE by Generality or TE but not by Generality. This corpus was built following the methodology described in Chapter 3.

Here we again followed the standard procedure to measure performance, namely we filled a confusion matrix with the number of true positive, false positive, false negative and true negative classifications produced by our system. From these we calculated various *Accuracy* and *Precision* scores.

5.3.1 All Words

On average over 8 AAMs, out of the 1203 positive TE by Generality pairs about 901 were correctly identified as such and the other about 302 were missed, i.e., our approach achieved 75% hit rate. Analyzing each measure individually, Table 5.14, we see that the measure that correctly classifies the greatest number *"Entailment by Generality"* pairs is the *J-measure*, with 997 correct, and the one with worst performance is *Braun-Blanket*, with 834 correct.

Table 5.14: Confusion Matrix for all AAM | All Words

AAM			System Response	
			A	*B*
Gold Standard	*Added Value*	*A*	978	441
		B	225	356
	Braun-Blanket	*A*	834	392
		B	369	405
	Certainty Factor	*A*	893	390
		B	310	407
	Conditional Probability	*A*	863	436
		B	340	361
	Conviction	*A*	893	286
		B	310	511
	Gini Index	*A*	891	394
		B	312	403
	J-measure	*A*	997	398
		B	206	399
	Laplace	*A*	856	383
		B	347	414

Table 5.15 shows that in terms of accuracy, for this approach, the best performing measures are the *Conviction* and the *J-Measure*. The *Conviction* shows the best performance in **Accuracy**, $0,7$, and in terms of **Precision for "Entailment, but no Generality"**, $0,64$. The *J-Measure* is the best measure in **Precision for "Entailment by Generality"**, $0,83$.

The worst results occur on measures *Added Value* and *Conditional Probability* in **Precision for "Entailment, but no Generality"**, both with $0,45$. The latter measure also shows a bad performance in terms of **Accuracy**, $0,61$. With respect to **Precision for "Entailment by Generality"** the measure with the worst performance is *Braun-Blanket*, $0,69$.

Table 5.15: Accuracy and Precision by AAM | All Words

AAM	Accuracy	Precision	
		A	*B*
ADDED VALUE	0,67	0,81	0,45
BRAUN-BLANKET	0,62	0,69	0,51
CERTAINTY FACTOR	0,65	0,74	0,51
CONDITIONAL PROBABILITY	0,61	0,72	0,45
CONVICTION	0,7	0,74	0,64
GINI INDEX	0,65	0,74	0,51
J-MEASURE	0,69	0,83	0,50
LAPLACE	0,64	0,71	0,52

It is worth noting, that the results on this subset are higher compared to the corresponding results presented in Section 5.2.1. Removing the cases that are not TE and dividing the rest in another pair of classes leads to this situation. A possible conclusion is that our methodology is better in telling TE by Generality from the other cases of TE than in the classification between TE and not an entailment. Similarly to the results in Section 5.2.1 the *J-Measure* achieves the best performance with respect to **Precision - Entailment** and also for **Precision for "Entailment by Generality"**.

5.3.2 Without Stop Words

On average, 862 TE by Generality pairs were correctly identified as such and the other 341 were classified as "Entailment but not Generality", i.e., this approach achieved 72% hit rate as Table 5.16 evidences. Analyzing measures individually we see that the measure that correctly classifies the greatest number *"Entailment by Generality"* pairs is the *Braun-Blanket*, with 965 correct. *J-measure* achieves comparable performance, classifying 943 correctly.

Table 5.16: Confusion Matrix for all AAM | Without Stop Words

AAM			System Response	
			A	*B*
Gold Standard	*Added Value*	*A*	889	421
		B	314	376
	Braun-Blanket	*A*	965	302
		B	238	495
	Certainty Factor	*A*	809	348
		B	394	449
	Conditional Probability	*A*	843	448
		B	360	349
	Conviction	*A*	756	383
		B	447	414
	Gini Index	*A*	861	388
		B	342	409
	J-measure	*A*	943	452
		B	260	345
	Laplace	*A*	833	394
		B	370	403

Table 5.17 shows that likewise the best performing measure in approach **"With All Words"**, in the present approach the best performing measure with respect to **Accuracy** and **Precision - Entailment by Generality** is *Braun-Blanket*. In this case, *Conviction* and *J-measure* are the measures with the worst performance.

Table 5.17: Accuracy and Precision by AAM | Without Stop Words

AAM	Accuracy	Precision	
		A	*B*
ADDED VALUE	0,63	0,74	0,47
BRAUN-BLANKET	0,73	0,80	0,62
CERTAINTY FACTOR	0,63	0,67	0,56
CONDITIONAL PROBABILITY	0,60	0,70	0,44
CONVICTION	0,59	0,63	0,52
GINI INDEX	0,64	0,72	0,51
J-MEASURE	0,64	0,78	0,43
LAPLACE	0,62	0,69	0,51

The behavior of the AAM in Table 5.17 is very different from the behavior presented in Section 5.2.2. On the corpus TE by Generality, in the approach **"With Stop Words"**, the *Braun-Blanket* has an excellent performance, the same is not the case for approach **"With All Words"**.

5.3.3 With Multiword Units

In this approach, only considering the subset of TE by Generality pairs, on average over the 8 AAM, $820\ T->H$ pairs where correctly classified and 383 pairs were missed, i.e., this approach achieves a hit rate of 68%, as shown in Table 5.18. However, in this approach we have two measures that stand out with an excellent performance. These are respectively *Added Value* and *Braun-Blanket*. We find that the best measure continues to be *Braun-Blanket*. This last measure classifies correctly 1113 *"Entailment by Generality"* pairs.

Table 5.18: Confusion Matrix for all AAM | With MWU

AAM			System Response	
			A	*B*
Gold Standard	*Added Value*	*A*	987	408
		B	216	389
	Braun-Blanket	*A*	1113	214
		B	90	583
	Certainty Factor	*A*	756	265
		B	447	532
	Conditional Probability	*A*	773	296
		B	430	501
	Conviction	*A*	674	396
		B	529	401
	Gini Index	*A*	786	233
		B	417	564
	J-measure	*A*	765	356
		B	438	441
	Laplace	*A*	703	296
		B	500	501

Table 5.19 shows that for this approach, the best performing AAM is the *Braun-Blanket* likewise the situation in Table 5.17, but here this measure has a greater emphasis.

In the same line of reasoning, when we look at Table 5.19, we find the excellent behavior that the measure *"Braun-Blanket"* has. For **"Accuracy"** we have $0,85$, for **"Precision - Entailment by Generality"** we have $0,93$, and for **"Precision - Entailment, but no Generality"** we have $0,73$. This result distinguishes *"Braun-Blanket"* from the other measures with a significant difference between **"Accuracy"** and **"Precision"**.

The worst performances figures are obtained by *Conviction* - $0,54$, *Laplace* - $0,58$ and *Added*

Value - $0,49$, (respectively **"Accuracy"**, **"Precision - Entailment by Generality"** and **"Precision - Entailment, but no Generality"**). These values here are lower compared to to previous approaches.

Table 5.19: Accuracy and Precision by AAM | All MWU

AAM	Accuracy	Precision	
		A	*B*
ADDED VALUE	0,69	0,82	0,49
BRAUN-BLANKET	0,85	0,93	0,73
CERTAINTY FACTOR	0,64	0,63	0,68
CONDITIONAL PROBABILITY	0,64	0,64	0,63
CONVICTION	0,54	0,56	0,50
GINI INDEX	0,68	0,65	0,71
J-MEASURE	0,60	0,64	0,55
LAPLACE	0,60	0,58	0,63

Although *Braun-Blanket* is the best measure **"With MWU"** approach on both the general TE set, Section 5.2.3, and on TE by Generality set, here the difference with the second best result is more pronounced.

5.3.4 Summary

In this section we summarize the results of the application of our methodology on the corpus TE by Generality. These are the results we are most interested in, as they concern the problem on which we are focusing our attention, namely identification of entailment by generality.

With respect to **Accuracy**, as seen in Table 5.20, the best performance, $0,85$, is achieved by the measure *Braun-Blanket* in conjunction with the approach *With MWU*. In this approach the second best measure is the *Added Value* whit $0,69$ of accuracy. We noted the significant difference between these two AAM.

The measure *Braun-Blanket* also is the best measure in approach *"Without Stop Words"*, with accuracy of $0,73$. In this approach, we have two measures with the second best performance, *Gini Index* and *J-measure*, with $0,64$ accuracy.

In *"All Words"*, we have two measures with the best performance, *Conviction* and *J-measure*, with $0,7$ and $0,69$ accuracy, respectively.

In Table 5.20, we realize that although *Conviction* is the best measure with *"All Words"* with respect to **Accuracy**, its performance is virtually equivalent to that of a random guesser *"Without Stop Words"* and *"With MWU"*.

Table 5.20: Accuracy by AAM

AAM	Accuracy		
	All Words	*Without Stop Words*	*With MWU*
ADDED VALUE	0,67	0,63	0,69
BRAUN-BLANKET	0,62	0,73	0,85
CERTAINTY FACTOR	0,65	0,63	0,64
CONDITIONAL PROBABILITY	0,61	0,60	0,64
CONVICTION	0,7	0,59	0,54
GINI INDEX	0,65	0,64	0,68
J-MEASURE	0,69	0,64	0,6
LAPLACE	0,64	0,62	0,6

With respect to **Precision**, the measure *Braun-Blanket* in conjunction with the approach *With MWU*, presents the best results in both **Precisions** *"Entailment by Generality"* (**A**) and **Precisions** *"Entailment, but no Generality"* (**B**), respectively $0,93$ and $0,73$.

For **Precisions** *"Entailment by Generality"*, the worst result is achieved by *Conviction* - $0,56$ *With MWU*, for **Precisions** *"Entailment, but no Generality"*, the worst result is achieved by *J-measure* - $0,43$ *Without Stop Words*.

Table 5.21: Precisions by AAM

AAM	Precision for A		
	All Words	*Without Stop Words*	*With MWU*
ADDED VALUE	0,81	0,74	0,82
BRAUN-BLANKET	0,69	**0,80**	**0,93**
CERTAINTY FACTOR	0,74	0,67	0,63
CONDITIONAL PROBABILITY	0,72	0,70	0,64
CONVICTION	0,74	0,63	0,56
GINI INDEX	0,74	0,72	0,65
J-MEASURE	**0,83**	0,78	0,64
LAPLACE	0,71	0,69	0,58
AAM	**Precision for B**		
	All Words	*Without Stop Words*	*With MWU*
ADDED VALUE	0,45	0,47	0,49
BRAUN-BLANKET	0,51	**0,62**	**0,73**
CERTAINTY FACTOR	0,51	0,56	0,68
CONDITIONAL PROBABILITY	0,45	0,44	0,63
CONVICTION	**0,64**	0,52	0,5
GINI INDEX	0,51	0,51	0,71
J-MEASURE	0,5	0,43	0,55
LAPLACE	0,52	0,51	0,63

5.4 Corpus TE by Generality translated into Portuguese

In this section we present the results of an experiment parallel to the one discussed in Section 5.3. Its original intention is to measure the degree to which our methodology is capable to recognize a specific kind of TE, namely TE by Generality. However, now we aim to study the possibility to adapt the processes to a different language. To this end we randomly selected a subset of 100 $T->H$ pairs from Corpus TE by Generality, preserving the proportion of 60 $T->H$ pairs of Entailment by Generality and 40 $T->H$ pairs of Entailment, but no Generality, and translated this subset into Portuguese using *Google Translate*[1] service.

Machine translation is a viable alternative to manual translation due to a combination of two factors. First, since our intention was to be as much language independent as possible, our methodology does not use morpho-syntactic analysis and language specific word order knowledge. On the other hand, *Google Translate* is reasonably successful in correct content word substitution. Thus, from

[1] `https://translate.google.pt/` [Last access: 21th December, 2013]

the perspective of our bag-of-words approach *Google Translate* preserves well the important information. This supposition is in line with the fact that our results in Portuguese are comparable to the corresponding results in English language.

5.4.1 All Words

On average over 8 AAMs, out of the 60 positive TE by Generality pairs about 44 were correctly identified as such and the other about 16 were missed, i.e., our approach achieved 73% hit rate. Analyzing each measure individually, we noticed that the Table 5.22 evidences a similar behavior to the Table 5.14 (even with the difference in the total number of pairs). In this approach, we see that the measure that correctly classifies the greatest number of *"Entailment by Generality"* pairs is the *J-measure*, with 51 correct, and the one with worst performance is *Braun-Blanket*, with 39 correct.

Table 5.22: Confusion Matrix for all AAM | All Words

AAM			System Response	
			A	*B*
Gold Standard	*Added Value*	*A*	47	24
		B	13	16
	Braun-Blanket	*A*	39	17
		B	21	23
	Certainty Factor	*A*	41	17
		B	19	23
	Conditional Probability	*A*	41	22
		B	19	18
	Conviction	*A*	44	16
		B	16	24
	Gini Index	*A*	43	17
		B	17	23
	J-measure	*A*	51	19
		B	9	21
	Laplace	*A*	42	21
		B	18	19

Table 5.23 shows that in terms of accuracy, for this approach, the best performing measures are the *J-Measure* and the *Conviction*. The *J-Measure* shows the best performance in **Accuracy**, $0,72$, and **Precision for "Entailment by Generality"**, $0,85$. The *Conviction* is the best measure in **Precision**

for "Entailment, but no Generality", $0,6$. As shown in Table 5.15, these measures also achieve good performance in English.

The worst result occur on measure *Added Value* in **Precision for "Entailment, but no Generality"**, with $0,4$. The measure *Conditional Probability* has bad performance in terms of **Accuracy**, $0,59$. With respect to **Precision for "Entailment by Generality"** the measure with the worst performance is *Braun-Blanket*, $0,65$.

Table 5.23: Accuracy and Precision by AAM | All Words

AAM	Accuracy	Precision	
		A	*B*
ADDED VALUE	0,63	0,78	0,4
BRAUN-BLANKET	0,62	0,65	0,58
CERTAINTY FACTOR	0,64	0,68	0,58
CONDITIONAL PROBABILITY	0,59	0,68	0,45
CONVICTION	0,68	0,73	0,6
GINI INDEX	0,66	0,72	0,58
J-MEASURE	0,72	0,85	0,53
LAPLACE	0,61	0,7	0,48

From Table 5.15 and Table 5.23 we see that the AAMs have a similar behavior in this approach, when compared over the English and Portuguese versions of the corpus. It is even more evident when we analyze the ranking of the measures in Table 5.15 and Table 5.23. That is, the AAM measures with the best performance in English are the same measures that show the best performance in Portuguese. Similarly, the worst measures in English also achieve low results in Portuguese.

5.4.2 Without Stop Words

On average over the 8 AAM measures 43 TE by Generality $T->H$ pairs were classified correctly and 17 were missed, i.e., this approach achieved a hit rate of about 72% as evidenced by Table 5.24. When the measures are considered individually, the ones that correctly classify the highest number of **"Entailment by Generality"** pairs are *Added Value* and *Braun-Blanket* with 47 correct, the *Gini Index* also has a good performance, classifying 45 pairs correctly.

Table 5.24: Confusion Matrix for all AAM | Without Stop Words

AAM			System Response	
			A	*B*
Gold Standard	*Added Value*	*A*	47	25
		B	13	15
	Braun-Blanket	*A*	47	16
		B	13	24
	Certainty Factor	*A*	39	17
		B	21	23
	Conditional Probability	*A*	39	22
		B	21	18
	Conviction	*A*	39	19
		B	21	21
	Gini Index	*A*	45	19
		B	15	21
	J-measure	*A*	43	25
		B	17	15
	Laplace	*A*	43	21
		B	17	19

Table 5.25 shows that the AAM with the best performance is the *Braun-Blanket*.

With respect to **Accuracy** and **Precision** for **"Entailment, but no Generality"** in the preset approach the best performing measure is *Braun-Blanket*. With respect to **Precision** for **"Entailment by Generality"** the measures *Braun-Blanket* and *Added Value* are the best performing measures, however, *Added Value* is the worst performing measure when **Precision** for **"Entailment, but no Generality"** is considered. *Conditional Probability* is the measure with the worst performance when compared by **Accuracy** criterion. The worst performers with respect to **Precision** for **"Entailment by Generality"** are *Certainty Factor*, *Conditional Probability* and *Conviction* measures.

Table 5.25: Accuracy and Precision by AAM | Without Stop Words

AAM	Accuracy	Precision	
		A	*B*
ADDED VALUE	0,62	0,78	0,38
BRAUN-BLANKET	0,71	0,78	0,6
CERTAINTY FACTOR	0,62	0,65	0,58
CONDITIONAL PROBABILITY	0,57	0,65	0,45
CONVICTION	0,6	0,65	0,52
GINI INDEX	0,66	0,75	0,53
J-MEASURE	0,58	0,72	0,38
LAPLACE	0,62	0,72	0,48

Braun-Blanket is the measure that shows the best performance both in English, Table 5.17, and in Portuguese, Table 5.25, when the corpus for TE by Generality is subjected to stop words removal (cf. Appendix 6.2).

5.4.3 With Multiword Units

In this approach, on average over the 8 AAM, 41 out of 60 $T->H$ pairs of TE by Generality were correctly classified as such and 19 pairs were missed, i.e., this approach achieved a hit rate of 68%, as we confirm in Table 5.26. In this approach we have two measures that stand out with an excellent performance, these are respectively *Added Value* and *Braun-Blanket*. The best measure continues to be *Braun-Blanket*. This last measure classifies correctly 53 pairs of *"Entailment by Generality"*.

Table 5.26: Confusion Matrix for all AAM | With MWU

AAM			System Response	
			A	*B*
Gold Standard	*Added Value*	*A*	51	22
		B	9	18
	Braun-Blanket	*A*	53	17
		B	7	23
	Certainty Factor	*A*	36	13
		B	24	27
	Conditional Probability	*A*	37	17
		B	23	23
	Conviction	*A*	33	23
		B	27	17
	Gini Index	*A*	41	13
		B	19	27
	J-measure	*A*	37	17
		B	23	23
	Laplace	*A*	36	13
		B	24	27

Table 5.27 shows that for this approach, the AAM with the best performance is *Braun-Blanket*, the same conclusion was drawn from the analysis of Table 5.19, however here this measure does not show significant advantage over the other measures. In the same line of reasoning, when we look at Table 5.27, we find the excellent behavior of the measure *"Braun-Blanket"* compared to the other measures, scoring $0,76$ with respect to **"Accuracy"**. *"Braun-Blanket"* also achieves the best result with respect to **"Precision - Entailment by Generality"**, $0,88$. *"Certainty Factor"*, *"Gini Index"* and *"Laplace"* achieve the highest **"Precision - Entailment, but no Generality"** value, $0,68$.
This approach achieves the absolute lowest values compared to the previous ones with *Conviction*, $0,5$, $0,55$ and $0,43$ (respectively **"Accuracy"**, **"Precision - Entailment by Generality"** and **"Precision - Entailment, but no Generality"**).

Table 5.27: Accuracy and Precision by AAM | All MWU

AAM	Accuracy	Precision	
		A	*B*
ADDED VALUE	0,69	0,85	0,45
BRAUN-BLANKET	0,76	0,88	0,58
CERTAINTY FACTOR	0,63	0,6	0,68
CONDITIONAL PROBABILITY	0,6	0,62	0,58
CONVICTION	0,5	0,55	0,43
GINI INDEX	0,68	0,68	0,68
J-MEASURE	0,6	0,62	0,58
LAPLACE	0,63	0,6	0,68

Comparing performance between English, Table 5.19, and Portuguese, Table 5.27, we confirm that there is no significant difference between both datasets. On the corpus TE by Generality translated into Portuguese, with the approach **"With MWU"** the *Braun-Blanket* has the best performance with respect to **"Accuracy"** and **"Precision - Entailment by Generality"** (likewise in Section 5.3.3). For **"Precision - Entailment, but no Generality"** the best AAMs are *"Certainty Factor"*, *"Gini Index"* and *"Laplace"*, unlike the corresponding results in Table 5.19.
In contrast to Section 5.3.3, here there is no a measure with significant advantage over the other measures.

5.4.4 Summary

In this section we summarize the results of the application of our methodology on the corpus TE by Generality translated into Portuguese.
With respect to **Accuracy** the best performance is achieved with the measure *Braun-Blanket* in conjunction with the approach **With MWU**, with result of $0,76$, as evidenced in Table 5.28. In this approach the second best measure is the *Added Value* whit result of $0,69$. Similarly, *Braun-Blanket* achieves the best performance in approach **"Without Stop Words"**, with result of $0,71$, followed by *Gini Index* with $0,66$. With **"All Words"**, the measure with the best **Accuracy** is *J-measure*, with $0,72$
From table 5.28, we read the three measures with the lowest **Accuracy**, namely, *Conditional Probability* with approach **"All Words"** and **"Without Stop Words"** and *Conviction* **"With MWU"**.

Table 5.28: Accuracy by AAM

AAM	Accuracy		
	All Words	*Without Stop Words*	*With MWU*
ADDED VALUE	0,63	0,62	0,69
BRAUN-BLANKET	0,62	0,71	0,76
CERTAINTY FACTOR	0,64	0,62	0,63
CONDITIONAL PROBABILITY	0,59	0,57	0,6
CONVICTION	0,68	0,6	0,5
GINI INDEX	0,66	0,66	0,68
J-MEASURE	0,72	0,58	0,6
LAPLACE	0,61	0,62	0,63

Considering the **Accuracy** figures for English and for Portuguese, presented in Table 5.20 and Table 5.28, which show similar scale and variations, we conclude that the performance of our methodology is not significantly influenced by the language.

With respect to **"Precision - Entailment by Generality"** the measure *Braun-Blanket* in conjunction with the approach *With MWU*, presents the best results of $0,88$, followed by the measure *J-measure* in conjunction with the approach *All Words*, $0,85$. The worst results are achieved in *With MWU* by *Certainty Factor* and *Laplace*, $0,6$.

With respect to **"Precision - Entailment, but no Generality"** the results are markedly lower. The best results are achieved in *With MWU* by *Certainty Factor*, *Gini Index* and *Laplace* with value of $0,68$. Moreover, the worst results are achieved by *Added Value*, $0,38$, in *All Words*.

Table 5.29: Precisions by AAM

AAM	Precision for A		
	All Words	*Without Stop Words*	*With MWU*
ADDED VALUE	0,78	**0,78**	0,85
BRAUN-BLANKET	0,65	**0,78**	**0,88**
CERTAINTY FACTOR	0,68	0,65	0,6
CONDITIONAL PROBABILITY	0,68	0,65	0,62
CONVICTION	0,73	0,65	0,55
GINI INDEX	0,72	0,75	0,68
J-MEASURE	**0,85**	0,72	0,62
LAPLACE	0,7	0,72	0,6
AAM	**Precision for B**		
	All Words	*Without Stop Words*	*With MWU*
ADDED VALUE	0,4	0,38	0,45
BRAUN-BLANKET	0,58	**0,6**	0,58
CERTAINTY FACTOR	0,58	0,58	**0,68**
CONDITIONAL PROBABILITY	0,45	0,45	0,58
CONVICTION	**0,6**	0,52	0,43
GINI INDEX	0,58	0,53	**0,68**
J-MEASURE	0,53	0,38	0,58
LAPLACE	0,48	0,48	**0,68**

Both, **Accuracy** and **Precision** figures show that whether applied to a corpus in English or in Portuguese language, our methodology provides classification capability significantly better than random guessing baseline and virtually indistinguishable with respect to the language.

5.5 Qualitative Analysis

In this chapter we study the behavior of our methodology for recognizing TE by Generality. Also, we provide a thorough comparison to relevant work. This is done taking under account the limitations of a typical language-independent and unsupervised learning techniques. In order to obtain fair comparison we used a well known dataset studied in the RTE Challenge as our test-bed. Further, as we are interested in a special kind of TE, we built a suitable corpus and also translated it into Portuguese language.

In this process we learned that detecting entailment between sentences is not an exact science.

We saw that each new RTE Challenge required different approach to the problem. Thus, we do not provide a measure or an approach that pretends to solve the problem. We can only conclude, based on evidences from Table 5.12 that for some combinations of measure and preprocessing approach our method shows good precision in recognizing TE.
Comparing our results, presented in Section 5.2, with the results of other relevant methodologies, presented in Section 2.2, we prove that our methodology achieves higher performance figures. In RTE-1 Challenge:

- The method described in Bayer *et al.* (2005), obtained $0,586$ of accuracy;
- The method described in Glickman & Dagan (2005), obtained $0,586$ of accuracy;
- The method described in Perez *et al.* (2005), obtained $0,495$ of accuracy.

On RTE-1 Challenge data, with approach *With MWU*, our methodology achieved better results than previous methodologies. Table 5.8 shows that the measures *Braun-Blanket* and *Laplace* achieve better results for **Weighted Average Accuracy**, namely $0,61$.
In the second case, the results are much more significant. As seen in Tables 5.20 and Table 5.21, there is always measure and an approach that stand out, namely *Braun-Blanket* measure **with MWU**. However *J-measure* and *Conviction* also have good results:

- *J-measure* in **Precision - Entailment by Generality** with **All Words**, has the second best performance (with $0,83$). In another words *J-measure* with **All Words** has a good performance to identify entailment by generality between sentences;
- *Conviction* ranks second for **Accuracy** (with $0,7$), and achieves a good result in *Precison - Entailment, but no generality or Other*, both with approach *All Words*.

Finally, when executed on the TE by Generality corpus translated in Portuguese, our methodology achieves results comparable to these in English language, although with less significant difference between the best and the following measures. However, in terms of **Accuracy**, Table 5.28, and **Precision - Entailment by Generality**, Table 5.29, the *Braun-Blanket* achieves the best performance in approach *With MWU*.
The results in both cases prove that there are several types of entailment (see Section 1.3.1), and evidence, through the corpus TE by Generality and the subset translated into Portuguese, that our method has a good performance in recognizing entailment by generality.

CHAPTER 6

CONCLUSION AND FUTURE WORK

"The best way to prepare for the future is to concentrate all the imagination and enthusiasm in executing the perfect job today."

Dale Carnegie

Finally, we present a recapitulation of this thesis and present our perspectives for the future investigation in this field.

6.1 Recapitulation

This Thesis presents and discusses the most relevant results of our research on TE by Generality. We studied the behavior of a specific variation of TE, where the objective is to understand how to identify entailment by generality between two sentences. We contribute a new direction for research in various fields of NLP.

In this thesis, we present an initial study of Entailment (Chapter 1), where we studied its meaning, its context in linguistics as well as its variations. This study contributed to promote a specific variation of Entailment - **Entailment by Generality** - and define the objective of our work - **Recognizing Textual Entailment by Generality**.

Still in Chapter 1, we show that there exists a strong relation between the concept of probability and TE. It is this relation that also supports our proposal to identify Entailment by Generality between sentences.

In Chapter 2 we present the study of the works that already exist in the area and survey all the approaches used in the first five RTE Challenges. Most of the participating systems relied on some external knowledge, e.g. linguistic tools (see Chapter 2), supervised approaches (see Chapter 2) and parameterized variables (for example threshold). In contrast, we presented a methodology for recognizing TE by Generality, one that is unsupervised, language-independent and threshold free solution.

In Chapter 2, we report the most important approaches used in RTE Challenges, with the variety of approaches, we realized that recognizing TE between two sentences is not an exact science. The datasets of RTE's Challenges consist of $T - H$ pairs with different levels of entailment reasoning,

based on lexical, syntactic, logical and world knowledge at different levels of difficulty. Through this chapter we show also the shortage of works that are language independent, unsupervised and threshold free in RTE Challenge.

Supported by the findings of the previous works and the results published in Pais *et al.* (2011) and Dias *et al.* (2011), we conclude that to evaluate the performance of our methodology, we could not just use the pairs of positive TE test datasets of the first five RTE Challenges. We needed a corpus of pairs of TE by Generality. Such a corpus did not exist previously.

In Chapter 3, we turn to a crowdsourcing platform called CrowdFlower in order to construct a corpus of pairs where TE by Generality relation holds. As a starting point of this exercise we used the set of the positive pairs from the datasets of the first five RTE Challenges. In this manner we obtained a manually annotated corpus, where the relationship between the text T and the hypothesis H of each pair is *Entailment by Generality*. This new corpus served us to study the behavior of our methodology.

With this corpus at hand, the next task (see Chapter 5) involves testing our methodology and evaluation of its performance in three different environments: i) All pairs of the Test Set of the first five RTE Challenges; ii) corpus of TE by Generality; and iii) a set of 100 $T - H$ pairs randomly taken from the corpus of TE by Generality and translated into Portuguese. To evaluate the performance of our methodology we used two benchmarks - the *Accuracy* (Equation 5.1) and the *Precision* (Equation 5.2). The results are very satisfactory and encouraging to further the work on our methodology.

In section 5.5 we looked at the quality of our results. We conclude that with respect to RTE by Generality, the measure *Braun-Blanket* in English and Portuguese, globally has an excellent performance, especially on approach *with MWU*. Also in this section, in conjunction with Section 5.2.4, we show that our methodology has better results on RTE-1 Challenge (in approach *With MWU*, see 5.2.3), than the other unsupervised and language-independent methodologies tested in this challenge (see Chapter 2.2).

With this thesis, we contribute an original proposal to RTE. Our methodology is unsupervised and language-independent, and accounts for the asymmetry of the studied phenomena by means of asymmetric similarity measures. However, we have demonstrated through the results in Chapter 5, that it is necessary to treat differently each type of textual entailment, i.e., through our methodology we got excellent results to identify pairs $T - H$ by generality, while, the result was less impressive when evaluating the data set of the first five RTE Challenges.

6.2 Future Research

Some of the issues addressed in this thesis give raise to interesting questions, problems and future research directions.

Our work does not end with the presentation of this thesis. This thesis is the beginning of our study in TE, more specifically TE by Generality. We will make available the corpus of TE by Generality to the scientific community. We intend to perform a syntactic and semantic analysis of this corpus in near future, in order to learn how many pairs each RTE Challenges contributed to the construction of this corpus, and finally what is the average size of T and H.

Still much work is to be done with the results presented in this thesis. It is necessary to make a more complete qualitative analysis, in order to understand what are the main differences between the various measures and approaches that we used. With the current results, we also want to know the measure and approach with more *Statistical Significance*((Demšar, 2006) (Foody, 2004)), independent of their performance. For this we will use the *McNemar's Test* and *ROC* analysis (Davis & Goadrich, 2006).

We want to improve our results and it is necessary to reinvent methodologies and improve existing ones, always respecting the idea of unsupervised, language-independent and threshold free solutions.

Thus, we are conducting further studies of the simplified asymmetric InfoSimba informative similarity measure $AISs(.\|.)$ (see Equation 4.13), which showed promising results (see Chapter 5). In this sense, the next task will be to implement our methodology in a new informative attributional similarity measure. To this end we turned the AIS into a N order similarity measure by proposing its recursive definition as in Equation 6.1, which we call the Recursive Asymmetric InfoSimba Similarity (RAIS), where the initialization is based on the initial version of the AIS i.e. $RAIS_0(X_i\|X_j) = AIS(X_i\|X_j)$. We also define its simplified version $RAISs_N(.\|.)$ in 6.2 with the following initialization $RAISs_0(X_i\|X_j) = AISs(X_i\|X_j)$.

$$RAIS_N(X_i\|X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.RAIS_{N-1}(W_{ik}\|W_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.RAIS_{N-1}(W_{ik}\|W_{il})+\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jl}.RAIS_{N-1}(W_{jk}\|W_{jl})-\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.RAIS_{N-1}(W_{ik}\|W_{jl})\end{array}\right)}. \tag{6.1}$$

$$RAISs_N(X_i\|X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.RAISs_{N-1}(W_{ik}\|W_{jl}). \tag{6.2}$$

With the wide usage of mobile devices summarizing Web pages *"on the fly"* is one of the most important applications for NLP. Following this direction, we plan to build a toolkit whose objective is to propose a summary of each generated cluster within the scope of ephemeral clustering (as mentioned in Section 1.2) search engine. Indeed, most of the time, the cluster label is not expressive enough to afford a clear understanding of the cluster content. For that purpose, we proposed an innovative solution, which is based on the discovery of the most expressive and general snippet within

a cluster based on the notion of TE. Our rationale is that the Web snippet, which best embodies a given cluster is the one, which entails all other ones with minimal loss of information. Once a user knows more about a cluster, she may find useful to understand what are the slight differences embodied by each Web page within the cluster. As such, each Web snippet could be highlighted (or ultra-summarized) by its differences and not its commonalities. These issues are very interesting for mobile information retrieval as well as for VIP users, as they may allow fast access to relevant information. Moreover, they can easily be computed in real-time based on our initial ideas.

References

ADAMS, R., NICOLAE, G., NICOLAE, C. & HARABAGIU, A. (2006). Textual entailment through extended lexical overlap. In *In Proceedings of RTE-2 Workshop*.

ADAMS, R., NICOLAE, G., NICOLAE, C. & HARABAGIU, S. (2007). Textual entailment through extended lexical overlap and lexico-semantic matching. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 119–124, Association for Computational Linguistics, Stroudsburg, PA, USA.

AGERRI, R. (2008). Metaphor in textual entailment. In D. Scott & H. Uszkoreit, eds., *COLING (Posters)*, 3–6.

ANDROUTSOPOULOS, I. & MALAKASIOTIS, P. (2010). A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.*, **38**, 135–187.

BAKER, C.F., FILLMORE, C.J. & LOWE, J.B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, 86–90, Association for Computational Linguistics, Stroudsburg, PA, USA.

BALAHUR, R., LLORET, E., FERRÁNDEZ, Ó., MONTOYO, A., PALOMAR, M. & MUÑOZ, R. (2008). The dlsiuaes team's participation in the tac 2008 tracks. In *In Proceedings of the Text Analysis Conference*.

BANKO, M. & BRILL, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, 26–33, Association for Computational Linguistics, Stroudsburg, PA, USA.

BARZILAY, R. & MCKEOWN, K.R. (2005). Sentence fusion for multidocument news summarization. *Comput. Linguist.*, **31**, 297–328.

BAYER, S., BURGER, J., FERRO, L., HENDERSON, J. & YEH, E. (2005). Mitre's submission to the eu pascal rte challenge. In *In PASCAL. Proc. of the First Challenge Workshop. Recognizing Textual Entailment*, 41–44.

BENSLEY, J. & HICKL, A. (2008). Workshop: Application of lcc's groundhog system for rte-4. In *In Proceedings of the Text Analysis Conference*.

BENTIVOGLI, L., DAGAN, I., DANG, H.T., GIAMPICCOLO, D. & MAGNINI, B. (2009). The fifth pascal recognizing textual entailment challenge. In *In Proc Text Analysis Conference (TAC'09*.

BOS, J. (2005). Combining shallow and deep nlp methods for recognizing textual entailment. In *In Proc. of the PASCAL RTE Challenge*, 65–68.

BOS, J. & MARKERT, K. (2006). When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*.

BOS, J. & OKA, T. (2007). A spoken language interface with a mobile robot. *Artificial Life and Robotics*, **11**, 42–47.

BROWN, P.F., PIETRA, V.J.D., PIETRA, V.J.D. & MERCER, R.L. (1991). Word-sense disambiguation using statistical methods. In *In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 264–270.

CALLISON-BURCH, C. & DREDZE, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, 1–12, Association for Computational Linguistics, Stroudsburg, PA, USA.

CANDELA, J.Q., DAGAN, I., MAGNINI, B. & BUC, F.D., eds. (2006). *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, vol. 3944 of *Lecture Notes in Computer Science*, Springer.

CARABALLO, S.A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, 120–126, Association for Computational Linguistics, Stroudsburg, PA, USA.

CHIERCHIA, G. & MCCONNELL-GINET, S. (2000). *Meaning and grammar (2nd ed.): an introduction to semantics*. MIT Press, Cambridge, MA, USA.

CHOUEKA, Y., KLEIN, T. & NEUWITZ, E. (1983). Automatic retrieval of frequent idiomatic and collocation expressions in a large corpus. *Journal for Literary and Linguistic Computing*, **4**, 34–38.

CLEUZIOU, G., DIAS, G. & LEVORATO, V. (2010). Modélisation prétopologique pour la structuration sémantico-lexicale. In *Proceedings of the 17èmes Rencontres de la Société Francophone de Classification (SFC 2010)*.

CLEUZIOU, G., BUSCALDI, D., LEVORATO, V. & DIAS, G. (2011). A pretopological framework for the automatic construction of lexical-semantic structures from texts. In C. Macdonald, I. Ounis & I. Ruthven, eds., *CIKM*, 2453–2456, ACM.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K. & TABLAN, V. (2002). Gate: an architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 168–175, Association for Computational Linguistics, Stroudsburg, PA, USA.

DAGAN, I. & GLICKMAN, O. (2004). Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining*.

DAGAN, I., MARCUS, S. & MARKOVITCH, S. (1993). Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, 164–171, Association for Computational Linguistics, Stroudsburg, PA, USA.

DAGAN, I., PEREIRA, F. & LEE, L. (1994). Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, 272–278, Association for Computational Linguistics, Stroudsburg, PA, USA.

DAGAN, I., GLICKMAN, O. & MAGNINI, B. (2005). The pascal recognising textual entailment challenge. In J.Q. Candela, I. Dagan, B. Magnini & F. d Alché-Buc, eds., *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, vol. 3944 of *Lecture Notes in Computer Science*, 177–190, Springer.

DAGAN, I., DOLAN, B., MAGNINI, B. & ROTH, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, **15**, i–xvii.

DAGAN, I., ROTH, D., SAMMONS, M. & ZANZOTTO, F.M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, **6**, 1–220.

DAVIS, J. & GOADRICH, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 233–240, ACM, New York, NY, USA.

DE BUENAGA, M., GACHET, D., MAÑA, M.J., DE LA VILLA, M. & MATA, J. (2008). Clustering and summarizing medical documents to improve mobile retrieval.

DELMONTE, R., TONELLI, S., BONIFORTI, A.P., BRISTOT, A. & PIANTA, E. (2005). Venses – a linguistically-based system for semantic evaluation. In *In Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment*, 49–52.

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.

DIAS, G. (2002). *Extraction Automatique d'Associations Lexicales à Partir de Corpora*. Ph.D. thesis, Univeristy of Orléans and New University of Lisbon.

DIAS, G. (2010). Information digestion. HDR Thesis, University of Oleans (France). 10 December. http://www.di.ubi.pt/~ddg/publications/Thesis-HDR.pdf.

DIAS, G., GUILLORÉ, S. & LOPES, J. (1999). Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of the 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, 333–339.

DIAS, G., ALVES, E. & LOPES, J. (2007). Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, 1334–1340.

DIAS, G., MUKELOV, R. & CLEUZIOU, G. (2008). Unsupervised graph-based discovery of general-specific noun relationships from web corpora frequency counts. In *Proceedings of the 12th International Conference on Natural Language Learning (CoNLL 2008)*.

DIAS, G., PAIS, S., WEGRZYN-WOLSKA, K. & MAHL, R. (2011). Recognizing textual entailment by generality using informative asymmetric measures and multiword unit identification to summarize ephemeral clusters. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, 284–287, IEEE Computer Society, Washington, DC, USA.

ESSEN, U. & STEINBISS, V. (1992). Cooccurrence smoothing for stochastic language modeling. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, ICASSP'92, 161–164, IEEE Computer Society, Washington, DC, USA.

FINKEL, J.R., GRENAGER, T. & MANNING, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 363–370, Association for Computational Linguistics, Stroudsburg, PA, USA.

FOODY, G.M. (2004). Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric engineering and remote sensing*, **70**, 627–634.

FREITAG, D., BLUME, M., BYRNES, J., CHOW, E., KAPADIA, S., ROHWER, R. & WANG, Z. (2005). New experiments in distributional representations of synonymy. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, 25–32.

GALE, W., CHURCH, K. & YAROWSKY, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, **26**, 415–439.

GIAMPICCOLO, D., MAGNINI, B., DAGAN, I. & DOLAN, B. (2007). The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 1–9, Association for Computational Linguistics, Stroudsburg, PA, USA.

GIAMPICCOLO, D., MAGNINI, B., DAGAN, I. & DOLAN, B. (2008). The fourth pascal recognizing textual entailment challenge. In *In Proceedings of the Text Analysis Conference*.

GLICKMAN, O. (2009). *APPLIED TEXTUAL ENTAILMENT: A generic framework to capture shallow semantic inference*. VDM Verlag, Saarbrücken, Germany, Germany.

GLICKMAN, O. & DAGAN, I. (2005). Web based probabilistic textual entailment. In *In Proceedings of the 1st Pascal Challenge Workshop*, 33–36.

GRISHMAN, R. & STERLING, J. (1993). Smoothing of automatically generated selectional constraints. In *Proceedings of the workshop on Human Language Technology*, HLT '93, 254–259, Association for Computational Linguistics, Stroudsburg, PA, USA.

GRISHMAN, R., HIRSCHMAN, L. & NHAN, N.T. (1986). Discovery procedures for sublanguage selectional patterns: initial experiments. *Comput. Linguist.*, **12**, 205–215.

GROSS, G. (1996). *Les Expressions Figées en Français*. Ophrys, Paris.

HARABAGIU, S. & HICKL, A. (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, 905–912, Association for Computational Linguistics, Stroudsburg, PA, USA.

HARABAGIU, S.M., MILLER, G.A. & MOLDOVAN, D.I. (1999). WordNet 2 – a morphologically and semantically enhanced resource. In *Proceedings of SigLex99: Standardizing Lexical Resources*, 1–8, University of Maryland.

HEARST, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, 539–545, Association for Computational Linguistics, Stroudsburg, PA, USA.

HERRERA, J., PEÑAS, A. & VERDEJO, F. (2005). Textual entailment recognition based on dependency analysis and wordnet. In *In Proceedings of the 1st. PASCAL Recognision Textual Entailment Challenge Workshop. Pattern Analysis, Statistical Modelling and Computational Learning*.

HICKL, A. & BENSLEY, J. (2007). A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 171–176, Association for Computational Linguistics, Stroudsburg, PA, USA.

HICKL, A., BENSLEY, J., WILLIAMS, J., ROBERTS, K., RINK, B. & SHI, Y. (2006). Recognizing textual entailment with lcc's groundhog system. In *In Proc. of the Second PASCAL Challenges Workshop*.

HUTCHINS, W.J., DOSTERT, L. & GARVIN, P. (1955). The georgetown-i.b.m. experiment. In *In*, 124–135, John Wiley & Sons.

IDO, R.B.H., DAGAN, I., DOLAN, B., FERRO, L., GIAMPICCOLO, D., MAGNINI, B. & SZPEKTOR, I. (2006). The second pascal recognising textual entailment challenge.

IFTENE, A. (2008). Uaic participation at rte4. In *In Text Analysis Conference (TAC 2008) Workshop - RTE-4 Track. National Institute of Standards and Technology (NIST*, 17–19.

IFTENE, A. & BALAHUR-DOBRESCU, A. (2007). Hypothesis transformation and semantic variability rules used in recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 125–130, Association for Computational Linguistics, Stroudsburg, PA, USA.

IFTENE, A. & MORUZ, M.A. (2009). UAIC Participation at RTE5. *Proceedings of TAC 2009*.

JIANG, J. & CONRATH, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, 19–33.

KAROV, Y. & EDELMAN, S. (1996). Learning similarity-based word sense disambiguation from sparse data.

KOUYLEKOV, M. & MAGNINI, B. (2005). Recognizing textual entailment with tree edit distance algorithms. In *PASCAL Challenges on RTE*, 17–20.

KULLBACK, S. & LEIBLER, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

LANDIS, J.R. & KOCH, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**, 159–174.

LI, F., ZHENG, Z., TANG, Y., BU, F., GE, R., ZHU, X. & ZHANG, X. (2008).

LI, F., ZHENG, Z., BU, F., TANG, Y., ZHU, X. & HUANG, M. (2009). Thu quanta at tac 2009 kbp and rte track. In *In Proc. of the Text Analysis Conference*.

LIN, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, 64–71, Association for Computational Linguistics, Stroudsburg, PA, USA.

LIN, D. (1998). Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

LIN, D. & PANTEL, P. (2001). Dirt - discovery of inference rules from text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 323–328.

LUND, K., BURGESS, C. & ATCHLEY, R. (1995). Semantic and associative priming in high dimensional semantic space. *Cognitive Science*, 660–665.

MANNING, C.D. & SCHÜTZE, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

MARCUS, M.P., MARCINKIEWICZ, M.A. & SANTORINI, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, **19**, 313–330.

MARTIN, W., AL, B. & VAN STERKENBURG, P. (1983). On the processing of a text corpus: From textual data to lexicographical information. In R. Hartman, ed., *Lexicography: Principles and Practice*, Applied Language Studies Series, Academic Press, London.

MEHDAD, Y., MOSCHITTI, R. & ZANZOTTO, F.M. (2009). Semker: Syntactic/semantic kernels for recognizing textual entailment. In *In Proc. of the Text Analysis Conference*.

MENÉNDEZ, M., PARDO, J., PARDO, L. & PARDO, M. (1997). The jensen-shannon divergence. *Journal of the Franklin Institute*, **334**, 307–318.

MICHELBACHER, L., EVERT, S. & SCHÜTZE, H. (2007). Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, 1–6.

MIHALCEA, R. & TARAU, P. (2004). Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

MILLER, G.A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, **38**, 39–41.

MILLER, G.A., LEACOCK, C., TENGI, R. & BUNKER, R.T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, 303–308, Association for Computational Linguistics, Stroudsburg, PA, USA.

MIRKIN, S., SPECIA, L., CANCEDDA, N., DAGAN, I., DYMETMAN, M. & SZPEKTOR, I. (2009). Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, 791–799, Association for Computational Linguistics, Stroudsburg, PA, USA.

NIELSEN, R.D., WARD, W. & MARTIN, J.H. (2009). Recognizing entailment in intelligent tutoring systems*. *Nat. Lang. Eng.*, **15**, 479–501.

OCH, F.J. & NEY, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, **29**, 19–51.

OHSHIMA, H. & TANAKA, K. (2009). Real time extraction of related terms by bi-directional lexico-syntactic patterns from the web. In *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC 2009)*, 441–449.

PADÓ, S., CER, D., GALLEY, M., JURAFSKY, D. & MANNING, C.D. (2009a). Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, **23**, 181–193.

PADÓ, S., GALLEY, M., JURAFSKY, D. & MANNING, C. (2009b). Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, 297–305, Association for Computational Linguistics, Stroudsburg, PA, USA.

PAIS, S., DIAS, G., WEGRZYN-WOLSKA, K., MAHL, R. & JOUVELOT, P. (2011). Textual entailment by generality. *Procedia - Social and Behavioral Sciences*, **27**, 258 – 266, computational Linguistics and Related Fields.

PALMER, M., GILDEA, D. & KINGSBURY, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, **31**, 71–106.

PAPINENI, K., ROUKOS, S., WARD, T. & ZHU, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318, Association for Computational Linguistics, Stroudsburg, PA, USA.

PAZIENZA, M.T. & PENNACCHIOTTI, M. (2005). Textual entailment as syntactic graph distance: a rule based and a svm based approach. In *In Proceedings PASCAL RTE challenge*, 528–535.

PAZIENZA, M.T., PENNACCHIOTTI, M. & ZANZOTTO, F.M. (2005). A linguistic inspection of textual entailment. In *Proceedings of the 9th Conference on Advances in Artificial Intelligence*, AI*IA'05, 315–326, Springer-Verlag, Berlin, Heidelberg.

PEÑAS, A., RODRIGO, A. & VERDEJO, F. (2008). Overview of the answer validation exercise 2007. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras & D. Santos, eds., *Advances in Multilingual and Multimodal Information Retrieval*, 237–248, Springer-Verlag, Berlin, Heidelberg.

PECINA, P. & SCHLESINGER, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006)*, 651–658.

PÉREZ, D. & ALFONSECA, E. (2006). Using bleu-like algorithms for the automatic recognition of entailment. In *Proceedings of the First international conference on Machine Learning Challenges: evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, 191–204, Springer-Verlag, Berlin, Heidelberg.

PEREZ, D., ALFONSECAIA, E. & RODRÍGUEZ, P. (2005). Application of the bleu algorithm for recognising textual entailments. In *Proceedings of the Recognising Textual Entailment Pascal Challenge*.

PUSTEJOVSKY, J., HANKS, P., SAURI, R., SEE, A., GAIZAUSKAS, R., SETZER, A., RADEV, D., SUNDHEIM, B., DAY, D., FERRO, L. & LAZO, M. (2003). The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, 647–656, Lancaster.

RESNIK, P. (1995). Disambiguating noun groupings with respect to wordnet senses. In *IN PROCEEDINGS OF THE THIRD WORKSHOP ON VERY LARGE CORPORA*, 54–68.

RILOFF, E. & SHEPHERD, J. (1997). A corpus-based approach for building semantic lexicons. In *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 117–124.

RODRIGO, A., PEÑAS, A. & VERDEJO, F. (2009). Overview of the answer validation exercise 2008. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, 296–313, Springer-Verlag, Berlin, Heidelberg.

ROMANO, L., KOUYLEKOV, M. & SZPEKTOR, I. (2006). Investigating a generic paraphrase-based approach for relation extraction. In *EACL '06*.

ROSCH, E. (1973). Natural categories. *Cognitive Psychology*, **4**, 265–283.

ROTH, D., SAMMONS, M. & VYDISWARAN, V. (2009). A framework for entailed relation recognition. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 57–60, Association for Computational Linguistics, Suntec, Singapore.

SAMMONS, M., VYDISWARAN, V.G.V., VIEIRA, T., JOHRI, N., CHANG, M.W., GOLDWASSER, D., SRIKUMAR, V., KUNDU, G., TU, Y., SMALL, K., RULE, J., QUANG, D. & ROTH, D. (2009). Relation alignment for textual entailment recognition. In *TAC*.

SANDERSON, M. & CROFT, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, 206–213, ACM, New York, NY, USA.

SANDERSON, M. & LAWRIE, D. (2000). Building, testing, and applying concept hierarchies. *Advances in Information Retrieval*, **7**, 235–266.

SCHANK, R.C. & TESLER, L. (1969). A conceptual dependency parser for natural language. In *Proceedings of the 1969 conference on Computational linguistics*, COLING '69, 1–3, Association for Computational Linguistics, Stroudsburg, PA, USA.

SCHÜTZE, H. (1992). Dimensions of meaning.

SCHÜTZE, H. (1993). Word space. In *Advances in Neural Information Processing Systems 5*, 895–902, Morgan Kaufmann.

SMADJA, F. (1993). Retrieving collocations from text: Xtract. *Comput. Linguist.*, **19**, 143–177.

SNOW, R., O'CONNOR, B., JURAFSKY, D. & NG, A.Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 254–263, Association for Computational Linguistics, Stroudsburg, PA, USA.

SZPEKTOR, I., TANEV, H., DAGAN, I. & COPPOLA, B. (2004). Scaling web-based acquisition of entailment relations. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona.

TAN, P.N., KUMAR, V. & SRIVASTAVA, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, **29**, 293–313.

TATU, M. & MOLDOVAN, D. (2007). Cogex at rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 22–27, Association for Computational Linguistics, Stroudsburg, PA, USA.

TATU, M., ILES, B., SLAVICK, J., NOVISCHI, A. & MOLDOVAN, D. (2006). Cogex at the second recognizing textual entailment challenge. In *In Proc. of the 2nd PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

WANG, R. & NEUMANN, G. (2007). Recognizing textual entailment using sentence similarity based on dependency tree skeletons. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 36–41, Association for Computational Linguistics, Stroudsburg, PA, USA.

WANG, R. & NEUMANN, G. (2008). An divide-and-conquer strategy for recognizing textual entailment. In *In Proc. of the Text Analysis Conference*.

WANG, R., ZHANG, Y. & NEUMANN, G. (2009). A joint syntactic-semantic representation for recognizing textual relatedness. In *Proceedings of TAC/RTE-5*.

WEIZENBAUM, J. (1966). Eliza - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, **9**, 36–45.

WOODS, W.A. (1970). Transition network grammars for natural language analysis. *Commun. ACM*, **13**, 591–606.

ZANZOTTO, F.M., MOSCHITTI, A., PENNACCHIOTTI, M. & PAZIENZA, M.T. (2006). Learning textual entailment from examples. In *In Proc. of the 2nd PASCAL Challenges Workshop on Recognising Textual Entailment*, 50–55.

ZHANG, C. & CHAI, J.Y. (2010). Towards conversation entailment: an empirical investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, 756–766, Association for Computational Linguistics, Stroudsburg, PA, USA.

Appendices

Appendix A

Stop Words Lists

A.1 Stop Words List in English

Table A.1: Stop Words List in English

Stop Words					
a	denne	estos	jenen	om	this
à	dennes	et	jener	on	ti
aber	der	être	jenes	or	til
af	dere	ett	jer	oss	to
affinché	deres	ettersom	kanskje	ou	tra
agli	desde	fazer	la	over	tu
ai	dess	for	lá	över	tua
al	dessa	för	laquelle	på	tuo
alla	det	fordi	le	para	tuoi
allo	dets	fra	lequel	pas	u
als	detta	från	les	pela	uden
an	deze	from	lo	pelo	uit
and	di	für	loro	perché	um
aquela	diese	gjøre	maar	pero	uma
aquele	diesem	gli	machen	por	un
aquello	diesen	göra	mais	porque	una
aquí	dieser	ha	más	qual	unas
aquilo	dieses	haben	me	que	und
är	disse	hacia	med	quegli	under
as	disses	han	mee	quei	une
at	dit	har	mi	quel	uno
até	ditt	här	mia	quella	unos
auf	do	have	mie	quelle	vad
aus	dopo	he	miei	quello	være
aux	dort	hebben	mig	questa	var
av	du	hebt	mio	queste	vara

Table A.1: (continued)

Stop Words					
avec	e	hennes	na	questi	våre
avere	è	hier	naar	questo	vars
b	een	hiermee	não	se	vid
bei	egli	hon	när	sé	vilka
but	ein	hos	negli	según	vilken
c	eine	hun	nei	sein	vilket
ce	einem	hur	nella	ser	você
ces	einen	hva	nelle	seu	vocês
cette	einer	hvem	nello	she	voi
che	eines	hvilke	ni	si	von
ci	el	hvilken	nicht	sie	voor
cioè	él	hvilket	niet	som	vosotros
come	ela	hvis	no	sondern	vostra
comme	elas	hvor	noi	sopra	vostre
como	ele	i	non	sotto	vostri
d	eles	ihr	nonché	su	vostro
da	ella	ikke	nós	sua	votre
dagli	ellas	il	nosotros	sugli	vous
dai	elle	ils	nosso	sui	we
dalla	eller	imidlertid	nostra	sul	welche
dallo	elles	in	nostre	sulla	welchem
där	ellos	innen	nostri	sulle	welchen
das	em	inte	nostro	sullo	welcher
dat	en	io	not	suo	welke
de	er	is	notre	suoi	wenn
deg	es	ist	nous	sur	which
degli	eso	isto	nu	te	wir
dei	essere	it	o	tener	with
della	essi	ja	och	ter	y
delle	esta	je	od	that	yo
dello	estas	jeg	oder	the	you
dem	este	jene	of	there	zijn
denna	esto	jenem	og	they	zu

A.2 Stop Words List in Portuguese

Table A.2: Stop Words List in Portuguese

Stop Words					
a	de	fernando	milhões	pode	sem
à	decisão	fez	ministério	podem	semana
acordo	depois	filme	ministro	poder	sempre
afirmou	desde	fim	momento	polícia	sendo
agora	desta	final	muito	política	ser
ainda	deste	foi	muitos	pontos	será
além	deve	folha	mundo	por	seria
algumas	dia	fora	música	porque	seu
alguns	dias	foram	na	porto	seus
ano	dinheiro	forma	nacional	portugal	sido
anos	direito	frente	nada	português	silva
antes	disse	governo	não	portuguesa	sistema
ao	diz	grande	nas	possível	situação
aos	dizer	grandes	nem	pouco	só
apenas	do	grupo	neste	preços	sobre
apesar	dois	guerra	no	presidente	social
após	dos	há	noite	primeira	sociedade
aqui	duas	história	nome	primeiro	sua
área	durante	hoje	nos	problema	suas
as	e	homem	nova	problemas	sul
às	é	início	novo	processo	tal
assim	economia	internacional	num	programa	também
até	ela	isso	numa	próprio	tão
através	ele	isto	número	próximo	tem
banco	eles	já	nunca	público	têm
bem	em	joão	o	qual	tempo
brasil	embora	jogo	obras	qualquer	ter
cada	empresa	josé	onde	quando	terá
câmara	empresas	lá	ontem	quanto	teve
capital	enquanto	lado	os	quase	tinha
carlos	então	lei	ou	quatro	toda
casa	entre	lhe	outra	que	todas
caso	era	lisboa	outras	quem	todo
cento	essa	livro	outro	quer	todos

Table A.2: (continued)

Stop Words					
central	esse	local	outros	questão	trabalho
centro	esta	lugar	país	r	três
cerca	está	maior	países	real	tudo
cidade	estado	maioria	para	região	último
cinco	estados	mais	parece	relação	últimos
coisa	estão	mas	parte	reportagem	um
com	estar	me	partido	república	uma
comissão	estava	meio	partir	rio	us
como	este	melhor	passado	são	vai
conta	eu	menos	paulo	saúde	valor
contos	eua	mercado	pela	se	vão
contra	exemplo	mês	pelas	segunda	ver
cultura	facto	meses	pelo	segundo	vez
da	falta	mesma	pelos	segurança	vezes
dar	faz	mesmo	pessoas	seis	vida
das	fazer	mil	plano	seja	zona

Appendix B

Multiword Units - Extraction of 2-ary Textual Associations

B.1 Multiword Units in English

Table B.1: MWU extracted from the first five RTE dataset test.

Mutual Expectation	Frequency	**MWU**
0.00152795552276	1791	of the
0.00082489388296	1276	in the
0.00063047156436	184	United States
0.00049295328790	136	Prime Minister
0.00024237485195	181	have been
0.00023694132688	198	has been
0.00022077299946	81	New York
0.00020527825109	86	took place
0.00019847218937	47	Los Angeles
0.00018817998352	587	to the
0.00018683138478	49	Saudi Arabia
0.00016676276573	348	is a
0.00014907041623	99	more than
0.00013466177916	52	Olympic Games
0.00012264038378	443	is the
0.00010284388554	112	will be
0.00009866995242	44	White House
0.00009571911505	36	prime minister
0.00009373646026	22	Victor Emmanuel
0.00009263328684	376	for the
0.00009158848115	34	Salvation Army
0.00008472465561	96	were killed
0.00008178089047	349	on the
0.00008129819616	25	Film Festival

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00007995819760	34	human rights
0.00007566171553	44	South Africa
0.00007240292325	18	Romano Prodi
0.00007196237857	17	al Qaeda
0.00007069162530	318	at the
0.00007027052197	278	in a
0.00006972133269	16	Vladislav Listyev
0.00006972133269	16	Celestial Seasonings
0.00006850526552	26	Yasser Arafat
0.00006850526552	26	Nelson Mandela
0.00006723128172	18	virtual reality
0.00006562007911	16	Port Nolloth
0.00006499827577	34	Bill Clinton
0.00006420765567	19	Salt Lake
0.00006338302774	20	la Cruz
0.00006197451876	16	Fiona Wood
0.00005871269968	16	Titanic sank
0.00005773863813	181	to be
0.00005693908679	14	Oberlin College
0.00005693908679	14	Condoleezza Rice
0.00005622687968	20	Tony Blair
0.00005602607052	15	Helmut Kohl
0.00005488655734	51	did not
0.00005460376269	44	European Union
0.00005455048813	26	Supreme Court
0.00005424439951	19	Philip Morris
0.00005339418567	22	Northern Ireland
0.00005338039409	14	Solar Temple
0.00005299763507	15	Dick Cheney
0.00005287018212	281	by the
0.00005229099770	12	Vasquez Rocks
0.00005161341323	47	years ago
0.00005052270353	20	Saddam Hussein
0.00005037366282	17	Neil Armstrong
0.00004887268369	69	had been

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00004880493361	14	Alfredo Cristiani
0.00004826861550	12	Wye Plantation
0.00004826861550	12	Jean Hackett
0.00004793341577	11	Chen Shui-bian
0.00004751171582	13	Qin Shi
0.00004694384552	260	from the
0.00004648088725	12	Standard Model
0.00004440148041	22	Latin America
0.00004361355968	34	oil prices
0.00004357583384	15	organizing committee
0.00004357583384	10	Li Zhaoxing
0.00004357583384	10	Aki Kaurismaki
0.00004357583384	10	Addis Ababa
0.00004327531133	12	Satellite Radio
0.00004221582276	39	United Nations
0.00004209590043	16	Stephen Harper
0.00004151313624	62	would be
0.00004145746425	54	million people
0.00004092339077	18	El Salvador
0.00004067077680	14	pleaded guilty
0.00004048335541	12	Harriet Lane
0.00003961439506	10	Nikos Kourkoulos
0.00003921824828	12	Arnold Schwarzenegger
0.00003921824828	9	Valerie Plame
0.00003914179979	16	per cent
0.00003789203038	10	Franz Liszt
0.00003766196824	11	Harry Potter
0.00003715413186	18	San Francisco
0.00003691129314	12	Nancy Pelosi
0.00003599015326	210	of a
0.00003529642345	9	Tommy Thompson
0.00003486066635	10	Silvio Berlusconi
0.00003486066635	10	Jack Ruby
0.00003486066635	8	XXIII Olympiade
0.00003486066635	8	Susan Linn

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00003486066635	8	Sharm el-Sheikh
0.00003486066635	8	Raman Raghav
0.00003486066635	8	Mo Siegel
0.00003486066635	8	Chadrick Fulks
0.00003486066635	8	Carrie Tomlinson
0.00003404583185	31	carried out
0.00003401726281	11	Nick Leeson
0.00003358244066	17	mercy killing
0.00003295422357	11	Real Madrid
0.00003281003956	8	Vicente Fox
0.00003281003956	8	serial killer
0.00003281003956	8	Moqtada al-Sadr
0.00003281003956	8	Charlton Heston
0.00003208765702	9	Stock Exchange
0.00003198226113	20	Security Council
0.00003172712240	18	White Sox
0.00003137459862	12	Eiffel Tower
0.00003137335443	213	with the
0.00003107488010	216	that the
0.00003098725938	8	True Path
0.00003098725938	8	Sierra Leone
0.00003098725938	8	Gertrude Jekyll
0.00003098725938	8	Elvis Presley
0.00003069254308	9	Kofi Annan
0.00003021579323	83	that it
0.00003005229883	10	Berni Ahern
0.00002976491305	25	World Cup
0.00002941368803	9	Ahmed Qurei
0.00002935634984	8	Gravetye Manor
0.00002935634984	8	fiber optic
0.00002905055590	10	scrap metal
0.00002905055590	10	Mahmoud Ahmadinejad
0.00002846112329	26	South Korea
0.00002788853271	8	Revenue Cutter
0.00002788853271	8	Heydar Aliyev

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00002728225991	12	interest rates
0.00002728225991	12	Baptist Church
0.00002724220394	124	as a
0.00002715109622	9	Hurricane Katrina
0.00002670178765	12	Stephen Hawking
0.00002621812564	76	that he
0.00002614549885	6	Ulan Bator
0.00002614549885	6	Radovan Karadzic
0.00002614549885	6	Niel Tupas
0.00002614549885	6	Costa Rica
0.00002614549885	6	conscientious objectors
0.00002614549885	6	Calista Flockhart
0.00002614549885	6	bronze bust
0.00002614549885	6	Brandenburg Gate
0.00002614549885	6	Anna Politkovskaya
0.00002614549885	6	Alison Hargreaves
0.00002590563236	18	so far
0.00002512018546	7	Christopher Reeve
0.00002490047518	10	Genie Awards
0.00002425089770	8	Benjamin Netanyahu
0.00002425089770	8	Andy Roddick
0.00002422075704	75	said that
0.00002413430775	12	El Nino
0.00002413430775	6	Tansu Ciller
0.00002413430775	6	Sri Lanka
0.00002413430775	6	petty thief
0.00002413430775	6	morning-after pill
0.00002413430775	6	Francis Ricciardone
0.00002388380744	53	such as
0.00002372868948	18	Gulf War
0.00002348508133	16	Winter Olympics
0.00002343411506	33	next year
0.00002330500138	19	North Korea
0.00002324474917	120	by a
0.00002321900502	38	this year

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00002247595512	7	Wolfgang von
0.00002247595512	7	Joseph Wilson
0.00002241042785	18	car bomb
0.00002241042785	6	Willy Claes
0.00002241042785	6	Javier Solana
0.00002241042785	6	Father Bátiz
0.00002241042785	6	Crathes castle
0.00002241042785	6	Aldrich Hazen
0.00002218406007	42	could be
0.00002178791692	5	Mabel Normand
0.00002178791692	5	Kurt Cobain
0.00002178791692	5	Industri Kapital
0.00002178791692	5	Buenos Aires
0.00002135215800	7	Panchen Lama
0.00002091639908	6	Umberto Bossi
0.00002091639908	6	Sonia Gandhi
0.00002091639908	6	Pedro Quintanar
0.00002091639908	6	Alberto Tomba
0.00002082231549	29	less than
0.00002075039629	10	rain forest
0.00002065817171	8	joint venture
0.00002065817171	8	Corfu Channel
0.00002063430111	47	at least
0.00001980719753	5	Satomi Mitarai
0.00001980719753	5	RJR Nabisco
0.00001980719753	5	Edvard Munch
0.00001980719753	5	ARENA assassins
0.00001960912414	6	shuttle Atlantis
0.00001960912414	6	Pamplona fiesta
0.00001960912414	6	Jessica Litman
0.00001960912414	6	Derek Plumbly
0.00001960912414	6	Audrey Seiler
0.00001923347190	8	Vladimir Meciar
0.00001923347190	8	Big Bang
0.00001856709423	7	lunar landing

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001845564657	12	Foreign Affairs
0.00001845564657	6	vision goggles
0.00001845564657	6	Princess Diana
0.00001845564657	6	long-term accommodation
0.00001815659743	10	arms embargo
0.00001815659743	5	Jesse Owens
0.00001743033317	4	Zack Urlocker
0.00001743033317	4	Yom Kippur
0.00001743033317	4	Viroj Laohaphan
0.00001743033317	4	Trevi Fountain
0.00001743033317	4	Tel Aviv
0.00001743033317	4	Sosnovyi Bor
0.00001743033317	4	Salvatore Gravano
0.00001743033317	4	Roger Federer
0.00001743033317	4	Reinventing Comics
0.00001743033317	4	Quentin Tarantino
0.00001743033317	4	Plaid Cymru
0.00001743033317	4	Padraig Pearse
0.00001743033317	4	Nicole Kidman
0.00001743033317	4	Merrill Lynch
0.00001743033317	4	Mein Kampf
0.00001743033317	4	Max Purnell
0.00001743033317	4	Markus Müller
0.00001743033317	4	Maggie Dempster
0.00001743033317	4	Lin Piao
0.00001743033317	4	Keith Maupin
0.00001743033317	4	Katamari Damacy
0.00001743033317	4	Josko Damic
0.00001743033317	4	Joachim Johansson
0.00001743033317	4	Ivan Getting
0.00001743033317	4	Irma Goldberg
0.00001743033317	4	Haggits Pillar
0.00001743033317	4	Goetz Friedrich
0.00001743033317	4	floating gardens
0.00001743033317	4	Clermont Ferrand

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001743033317	4	Budleigh Salterton
0.00001743033317	4	Borislav Shervinsky
0.00001743033317	4	Aurore Paquiss
0.00001743033317	4	Andre Agassi
0.00001743033317	4	Alaattin Cakici
0.00001743033317	4	Agua Dulce
0.00001708172567	7	gold medals
0.00001708172567	7	Colin Powell
0.00001660685302	22	set up
0.00001651294770	6	Secret Service
0.00001651294770	6	Dalai Lama
0.00001651294770	6	bomber detonated
0.00001641694143	9	auction houses
0.00001640501978	8	Tata Steel
0.00001640501978	8	Gerhard Schroeder
0.00001631775922	44	known as
0.00001619975774	31	President Bush
0.00001593630441	8	William Doyle
0.00001591751243	22	most important
0.00001588339182	27	do not
0.00001587777479	94	at a
0.00001568729931	6	Le Boucher
0.00001568729931	6	Ice Hockey
0.00001568729931	6	fishing quotas
0.00001556279676	5	Aptis Communications
0.00001549362969	4	Symbian OS
0.00001549362969	4	Steve Price
0.00001549362969	4	Sergey Brin
0.00001549362969	4	Sergei Sidorsky
0.00001549362969	4	Jeff Hunt
0.00001549362969	4	Ibrahim Sofu
0.00001549362969	4	Cate Blanchett
0.00001534627154	9	Green Zone
0.00001531962880	15	presidential election
0.00001508394234	30	last year

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001497346329	35	can be
0.00001480157334	25	does not
0.00001452527795	5	Margaret Thatcher
0.00001447458817	91	with a
0.00001426118160	6	Super Bowl
0.00001426118160	6	Mount Graham
0.00001394426636	4	Vue Cinemas
0.00001394426636	4	Todd Keys
0.00001394426636	4	terminally ill
0.00001394426636	4	Soto Toro
0.00001394426636	4	Perez Quiron
0.00001394426636	4	Pan Am
0.00001394426636	4	non-manufacturing index
0.00001394426636	4	Mohamed Fulayfel
0.00001394426636	4	Kailash Satyarthi
0.00001394426636	4	infinite canvas
0.00001394426636	4	Genetic Modification
0.00001394426636	4	Ciudad Madero
0.00001394426636	4	Buzz Aldrin
0.00001394426636	4	Berry Gordy
0.00001377558601	7	mental illness
0.00001366538163	14	postmenopausal women
0.00001361744762	5	Dow Jones
0.00001318616614	94	was a
0.00001307274943	3	Rossville Blvd
0.00001307274943	3	Roh Moo-hyun
0.00001307274943	3	Palos Verdes
0.00001307274943	3	Notre Dame
0.00001307274943	3	MG Rover
0.00001307274943	3	Lleyton Hewitt
0.00001307274943	3	Hong Kong
0.00001307274943	3	Hector Oqueli
0.00001307274943	3	Gustavo Guzman
0.00001307274943	3	Enola Gay
0.00001307274943	3	asylum seekers

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001307274943	3	Alexandre Berthier
0.00001297768631	92	for a
0.00001283506390	9	Air Force
0.00001267660627	8	industrial espionage
0.00001267660627	4	Writer Samir
0.00001267660627	4	Sony BMG
0.00001267660627	4	parking garage
0.00001267660627	4	Judith Miller
0.00001267660627	4	Houston Rockets
0.00001256009273	7	Wall Street
0.00001254983999	6	Katharine Hepburn
0.00001237700144	13	high school
0.00001226203676	11	Central Bank
0.00001212544885	8	Filipino hostage
0.00001210439768	5	Time Warner
0.00001199506823	16	New Zealand
0.00001190198327	105	and a
0.00001162022181	4	Walter Cronkite
0.00001162022181	4	Martian meteorite
0.00001162022181	4	Marilyn Manson
0.00001162022181	4	Larry Page
0.00001162022181	4	Grand Prix
0.00001162022181	4	CBS newsman
0.00001162022181	4	Andrei Kozyrev
0.00001162022181	4	Andrei Kokoshin
0.00001162022181	4	Achille Occhetto
0.00001146732484	5	Rafik Hariri
0.00001146732484	5	Giulio Andreotti
0.00001146732484	5	Dave McCool
0.00001139884171	29	when he
0.00001138781772	14	no longer
0.00001138594325	9	unemployment rate
0.00001138307471	8	De Klerk
0.00001120521392	3	Viktor Yanukovych
0.00001120521392	3	Las Vegas

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001120521392	3	Hugo Chávez
0.00001120521392	3	Gilda Flores
0.00001108490596	103	to a
0.00001082987637	13	foreign minister
0.00001077239449	24	three years
0.00001072635860	4	Valdez Principles
0.00001072635860	4	Moroccan delegation
0.00001072635860	4	modem connections
0.00001071536917	15	security forces
0.00001067607900	7	mobile phones
0.00001063322907	121	after the
0.00001054608401	12	news conference
0.00001031532793	19	last week
0.00001028176757	52	he was
0.00001012083885	6	intelligence agencies
0.00000996019025	12	opposition parties
0.00000996019025	4	Metin Kaplan
0.00000996019025	4	Mel Sembler
0.00000996019025	4	Mel Karmazin
0.00000996019025	4	measuring 6.9
0.00000996019025	4	Jurgen Schneider
0.00000996019025	4	browser bug
0.00000996019025	4	Ayatollah Ali
0.00000990359877	5	particle physics
0.00000988811917	28	may be
0.00000988683041	78	on a
0.00000980456207	6	Social Democrats
0.00000980456207	3	vacuum tubes
0.00000980456207	3	Soledad Prison
0.00000980456207	3	Piazza dei
0.00000980456207	3	Federico Fellini
0.00000980456207	3	en route
0.00000968011955	118	as the
0.00000953957442	9	Nobel Prize
0.00000929617727	4	Lawrence Welk

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000929617727	4	advisory panel
0.00000896417168	12	member states
0.00000896417168	6	intensive care
0.00000871516659	4	Yitzhak Rabin
0.00000871516659	4	Ohama Beach
0.00000871516659	4	minke whales
0.00000871516659	4	Korkmaz Yigit
0.00000871516659	4	Klaus Wowereit
0.00000871516659	4	Klaus Pohl
0.00000871516659	4	Frank Costello
0.00000871516659	3	São Paulo
0.00000871516659	3	negative publicity
0.00000871516659	2	Zeev Bielsky
0.00000871516659	2	Yevgenia Timoshenko
0.00000871516659	2	Wu Yi
0.00000871516659	2	Verdens Gang
0.00000871516659	2	Uttar Pradesh
0.00000871516659	2	Sylvia Costas
0.00000871516659	2	Silicon Graphics
0.00000871516659	2	Shapour Bakhtiar
0.00000871516659	2	Salvo Lima
0.00000871516659	2	Sally Struthers
0.00000871516659	2	Ronnie Gilbert
0.00000871516659	2	Regina Schueller
0.00000871516659	2	Olive Madison
0.00000871516659	2	Nicol Williamson
0.00000871516659	2	Natwar Singh
0.00000871516659	2	Muktar Said-Ibrahim
0.00000871516659	2	Morihiro Hosokawa
0.00000871516659	2	Minnie Driver
0.00000871516659	2	Minerva Rigging
0.00000871516659	2	milepost markers
0.00000871516659	2	Midge Ure
0.00000871516659	2	Michel Camdessus
0.00000871516659	2	Melvyn Percy

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000871516659	2	Malcolm Brabant
0.00000871516659	2	Lindsay Ellis
0.00000871516659	2	Les Combes
0.00000871516659	2	Kiryat Arba
0.00000871516659	2	Kaspersky Labs
0.00000871516659	2	Karl Rove
0.00000871516659	2	Jubiz Hazvumb
0.00000871516659	2	Henrik Larsson
0.00000871516659	2	Helicobacter pylori
0.00000871516659	2	Guillermo Ortiz
0.00000871516659	2	Gene Bartow
0.00000871516659	2	Forza Italia
0.00000871516659	2	first-time offenders
0.00000871516659	2	Duane Hanson
0.00000871516659	2	diplomacy directives
0.00000871516659	2	Cyprian Gatete
0.00000871516659	2	Contemnit procellas
0.00000871516659	2	Clement VII
0.00000871516659	2	Chuck Yeager
0.00000871516659	2	Celso Amorim
0.00000871516659	2	Camille Pissarro
0.00000871516659	2	Bracamonte Battalion
0.00000871516659	2	Berliner Verlag
0.00000871516659	2	Ben Stiller
0.00000871516659	2	Beatriz Iero
0.00000871516659	2	Baz Luhrman
0.00000871516659	2	Barrow Hanley
0.00000871516659	2	Ayrton Senna
0.00000871516659	2	Angela Merkel
0.00000871516659	2	Andriy Shevchenko
0.00000871516659	2	alter ego
0.00000871516659	2	African-American WWE
0.00000869458745	65	a new
0.00000820250989	4	Kofi Anan
0.00000820250989	4	Financial Times

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000820250989	4	attorney fees
0.00000784364966	3	Samir Geagea
0.00000784364966	3	Jacques Delors
0.00000784364966	3	Houston multimillionaire
0.00000784364966	3	hot springs
0.00000784364966	3	Genghis Khan
0.00000774681484	4	Bosnian Serb
0.00000774681484	4	beauty contests
0.00000764155902	11	six months
0.00000756014470	12	Mr Putin
0.00000751307471	5	Alfred Hitchcock
0.00000733908746	8	health care
0.00000733908746	4	Tikrit my
0.00000733908746	4	stores fund
0.00000733908746	4	Saint Laurent
0.00000733908746	4	poli outbreak
0.00000733908746	4	Eugene Shoemaker
0.00000733908746	4	eighteen books
0.00000733908746	4	23 rd
0.00000713059080	3	monetary union
0.00000713059080	3	Maurice Strong
0.00000713059080	3	Linux LTS
0.00000713059080	3	Don Brash
0.00000697213318	8	best picture
0.00000697213318	4	tape containing
0.00000697213318	4	holiday weekend
0.00000697213318	2	Stefano Falconi
0.00000697213318	2	spaceship launches
0.00000697213318	2	haemorrhagic fever
0.00000697213318	2	Gerald Schatten
0.00000697213318	2	genetically modified
0.00000697213318	2	Deng Xiaoping
0.00000697213318	2	Aurora Borealis
0.00000697213318	2	800 Chechens
0.00000697213318	2	357 yards

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000688605769	8	growth rate
0.00000680872381	5	tropical storm
0.00000664405570	74	of its
0.00000664012668	4	Warner Village
0.00000664012668	4	Van Hasselt
0.00000664012668	4	Channel Tunnel
0.00000664012668	4	ballistic missile
0.00000659084481	11	war crimes
0.00000653637471	3	Fidel Castro
0.00000647035085	7	border dispute
0.00000642330406	93	during the
0.00000633830314	4	shallow waters
0.00000633830314	4	Mahmoud Ahmadi-Nejad
0.00000633830314	4	Falkland islanders
0.00000620589526	28	they are
0.00000611031555	21	should be
0.00000606272442	4	Nick Berg
0.00000606272442	4	Clark Gable
0.00000603357694	6	natural gas
0.00000581011091	2	Virtual Reality
0.00000581011091	2	Robbie Williams
0.00000581011091	2	Rick Dinon
0.00000581011091	2	Madhu Mani
0.00000581011091	2	Judy Sgro
0.00000581011091	2	Jake Peavy
0.00000581011091	2	Harnold Lamb
0.00000581011091	2	Gianni Versace
0.00000581011091	2	Ghazi al-Yawar
0.00000581011091	2	fastest swimmer
0.00000581011091	2	Doug Arthur
0.00000581011091	2	Bolan Pass
0.00000581011091	2	Balinese massage
0.00000581011091	2	Alabama Birmingham
0.00000569390886	7	both sides
0.00000566687550	53	and other

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000558664533	5	Ryder Cup
0.00000537707729	25	said it
0.00000531210162	8	second round
0.00000508384710	7	heart disease
0.00000498009513	6	steel manufacturer
0.00000498009513	2	Tennessee Tech
0.00000498009513	2	Iron Curtain
0.00000473047794	80	the US
0.00000466095935	13	2 million
0.00000461391164	6	King Abdullah
0.00000435758329	3	polling stations
0.00000435758329	2	Ross Perot
0.00000435758329	2	Jake Borski
0.00000435758329	2	conjugated estrogen
0.00000434792128	30	as an
0.00000407876178	32	for its
0.00000407049356	48	from a
0.00000387340742	2	Robin Warren
0.00000387340742	2	Renzo Piano
0.00000387340742	2	Raisani tribe
0.00000387340742	2	Rafael Benitez
0.00000387340742	2	patent abstract
0.00000387340742	2	Miss Universe
0.00000387340742	2	Mauricio Pineda
0.00000387340742	2	Lorenzo brothers
0.00000387340742	2	Liverpool boss
0.00000387340742	2	Kristina Miller
0.00000387340742	2	Douglas Hurd
0.00000387340742	2	Deutsche Telekom
0.00000387340742	2	crushing debt
0.00000387340742	2	bright ideas
0.00000387121281	28	works for
0.00000385592011	14	when they
0.00000374342721	8	officials say
0.00000363440995	14	but also

Table B.1: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000361625166	70	over the
0.00000358648822	10	three months
0.00000353019414	8	returned home
0.00000348606659	2	Sue Robbins
0.00000348606659	2	Mary Vattimo
0.00000348606659	2	Jim Lankes
0.00000348606659	2	burning backpack
0.00000348606659	2	Brian Goodell
0.00000348606659	2	Bob Geldof
0.00000348606659	2	Allan Chapman
0.00000348606659	2	221 BCE
0.00000342776184	54	of his
0.00000332398781	44	to his
0.00000316915157	2	replacement therapy
0.00000316915157	2	precious cargo
0.00000316915157	2	Ian Woosnam
0.00000316915157	2	green card
0.00000316915157	2	genetic illnesses
0.00000316915157	2	Christopher Yavelow
0.00000306932088	53	is in
0.00000290505545	3	binge drinking
0.00000270470696	3	Fair Trade
0.00000256328440	5	presidential candidate
0.00000253020971	3	maternity leave
0.00000245114052	3	tax cuts
0.00000230695582	3	Apple Computer

B.2 Multiword Units in Portuguese

Table B.2: MWU extracted from the first five RTE dataset test, translated into Portuguese.

Mutual Expectation	Frequency	**MWU**
0.00080304709263	242	Estados Unidos
0.00039809668669	117	Reino Unido
0.00035627873149	129	US $
0.00025095048477	80	Jogos Olímpicos
0.00020497686637	514	para o
0.00019914630684	59	Los Angeles
0.00019199564122	519	para a
0.00018472748343	203	dos EUA
0.00015711436572	46	Nações Unidas
0.00015202724899	358	é um
0.00014415156329	72	Nova York
0.00013839868188	443	que o
0.00013588291768	328	disse que
0.00012119788880	353	com o
0.00011721674673	388	em um
0.00011596808326	36	Casa Branca
0.00011264761997	342	é o
0.00011245389760	48	Barack Obama
0.00009970800602	30	Arábia Saudita
0.00008909270400	319	com a
0.00008661813627	308	em uma
0.00008506343147	361	que a
0.00008297820750	253	que ele
0.00007896460011	460	de um
0.00007609972090	26	Relações Exteriores
0.00006711115566	25	companhia aérea
0.00006578615285	130	por cento
0.00006496359856	22	Brian Mulroney
0.00006482122262	41	União Europeia
0.00005798404163	36	direitos humanos
0.00005569440691	56	ano passado
0.00005565052561	67	ter sido
0.00005507421156	68	pelo menos
0.00005319786942	25	San Francisco

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00005234670243	15	Hong Kong
0.00005139494533	18	Condoleezza Rice
0.00005065809819	15	Las Vegas
0.00004985400301	20	realidade virtual
0.00004758791329	15	Negócios Estrangeiros
0.00004743390673	70	tem sido
0.00004717220145	14	Second Life
0.00004594009079	21	Nelson Mandela
0.00004558844012	205	que os
0.00004536714187	13	Sri Lanka
0.00004419572724	170	é uma
0.00004364085908	186	para os
0.00004222634016	22	Manchester United
0.00004084039756	32	muitas vezes
0.00004009382610	316	de uma
0.00003955084321	17	Gordon Brown
0.00003865602775	12	Romano Prodi
0.00003838758130	11	Buenos Aires
0.00003758734601	63	foram mortos
0.00003738382657	100	todos os
0.00003665222539	31	Nova Zelândia
0.00003489780283	15	Oriente Médio
0.00003489780283	10	Meio Ambiente
0.00003489780283	10	Falun Gong
0.00003489780283	10	Aki Kaurismaki
0.00003465712871	12	Sam Brownback
0.00003380602720	51	tinha sido
0.00003371259299	16	Saddam Hussein
0.00003297430158	180	a sua
0.00003248668145	16	Yasser Arafat
0.00003204899986	60	pode ser
0.00003140802073	12	Torre Eiffel
0.00003140802073	9	Bento XVI
0.00003136881060	20	seres humanos
0.00003072708932	19	Supremo Tribunal

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00003072708932	19	nave espacial
0.00002871590550	12	White Sox
0.00002840976049	17	Vladimir Putin
0.00002791824227	8	Thaksin Shinawatra
0.00002776360088	151	sobre o
0.00002684446372	10	Arnold Schwarzenegger
0.00002661696817	15	El Salvador
0.00002644886081	12	carne bovina
0.00002561673682	64	mais tarde
0.00002536657485	26	New York
0.00002481621414	8	Angelina Jolie
0.00002481621414	8	Creative Commons
0.00002442846198	7	Alfredo Cristiani
0.00002442846198	7	Yigal Amir
0.00002405763189	130	como um
0.00002389926340	140	que as
0.00002380397382	18	ficaram feridos
0.00002359091377	13	Prêmio Nobel
0.00002355601646	9	Força Aérea
0.00002351009789	8	empréstimos ruins
0.00002351009789	8	Andy Roddick
0.00002337341175	12	Harry Potter
0.00002268357093	13	Tony Blair
0.00002251471051	10	Movimento Esquerda-Verde
0.00002251471051	10	Ron Gray
0.00002244378629	39	estão sendo
0.00002233459236	8	Carla Bruni
0.00002233459236	12	Suprema Corte
0.00002229975871	126	um dos
0.00002222438889	11	homicídio culposo
0.00002222438889	11	Rainha Elizabeth
0.00002192352622	32	nos últimos
0.00002184415098	137	o seu
0.00002174401561	18	New Line
0.00002166327795	158	é a

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00002135272189	19	Hillary Clinton
0.00002127104199	8	Vicente Fox
0.00002115018287	20	Bill Clinton
0.00002093868170	9	Mahmoud Ahmadinejad
0.00002093868170	6	Celso Amorim
0.00002093868170	6	Ramos Horta
0.00002093868170	6	Datuk Seri
0.00002093868170	6	Umberto Bossi
0.00002081797174	128	por um
0.00002069378388	13	São Paulo
0.00002033699457	13	ficaram feridas
0.00002032407974	114	do governo
0.00002030417636	8	Opera House
0.00002019087151	9	Nicolas Sarkozy
0.00002011755714	7	Scotland Yard
0.00001988475742	43	- year-old
0.00001949463331	9	gripe suína
0.00001942138624	8	Nancy Pelosi
0.00001938766763	10	Red Bull
0.00001932801388	6	Wye Plantation
0.00001932801388	6	Angela Merkel
0.00001932801388	6	Johanna Sigurdardottir
0.00001932801388	6	Garcia Marquez
0.00001932801388	6	Victor Emmanuel
0.00001913980850	111	com uma
0.00001899991366	7	Anna Politkovskaya
0.00001890897693	38	à noite
0.00001886367681	10	la Cruz
0.00001876820716	21	ataque cardíaco
0.00001861216151	8	Motor Company
0.00001823691491	9	Zona Verde
0.00001813572089	151	do que
0.00001799991878	7	Dick Cheney
0.00001799991878	7	Wen Jiabao
0.00001799991878	7	tosse convulsa

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001794744094	6	Li Zhaoxing
0.00001794744094	6	Valentino Rossi
0.00001794744094	6	Cyril Ferez
0.00001794744094	6	Javier Solana
0.00001794744094	6	Michel Fourniret
0.00001794744094	6	Benazir Bhutto
0.00001786767461	8	telefones móveis
0.00001766596870	46	está sendo
0.00001744890142	5	Sharm el-Sheikh
0.00001744890142	5	Andry Rajoelina
0.00001744890142	5	Dobie Gillis
0.00001744890142	5	Helmut Kohl
0.00001744890142	5	Beni Suef
0.00001709992284	7	Porto Nolloth
0.00001709992284	7	Oberlin College
0.00001709992284	7	Yitzhak Rabin
0.00001700284702	119	com um
0.00001675094427	6	Cessna 172
0.00001675094427	6	Jefferson Airplane
0.00001628564132	14	candidato presidencial
0.00001628564132	7	Manmohan Singh
0.00001615269684	9	Real Madrid
0.00001586263716	10	ferimentos graves
0.00001586263716	5	Temperos Celestial
0.00001586263716	5	Nikos Kourkoulos
0.00001586263716	10	Air Canada
0.00001570401037	6	Britney Spears
0.00001570401037	6	Awatef Aboudihaj
0.00001565375896	121	sobre a
0.00001556631832	91	com os
0.00001554538358	7	Fiona Wood
0.00001551013338	10	Walt Disney
0.00001512133895	110	o governo
0.00001494508069	48	todas as
0.00001488972885	8	Direitos Humanos

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001487960617	56	as pessoas
0.00001486949805	7	Silvio Berlusconi
0.00001486949805	7	Time Warner
0.00001486949805	7	Singapore Airlines
0.00001486949805	7	Chen Shui-bian
0.00001482870539	36	deve ser
0.00001478024569	6	Harriet Lane
0.00001462297041	49	as autoridades
0.00001456603877	12	República Checa
0.00001454075118	5	fast food
0.00001440941469	16	sistema operacional
0.00001424993570	7	insuficiência cardíaca
0.00001413360951	9	Al Qaeda
0.00001397829510	27	este ano
0.00001395912113	4	XXIII Olympiade
0.00001395912113	4	Adis Abeba
0.00001395912113	4	Chadrick Fulks
0.00001395912113	4	joint venture
0.00001395912113	4	Valerie Plame
0.00001395912113	4	Katamari Damacy
0.00001395912113	4	Edouard Balladur
0.00001395912113	4	Dalai Lama
0.00001395912113	4	Yad Vashem
0.00001395912113	4	Nadia Comaneci
0.00001395912113	4	Ocean Drive
0.00001395912113	4	Luciano Bello
0.00001395912113	4	Pita Sharples
0.00001395912113	4	Desejo Tagro
0.00001395912113	4	Vlaams Belang
0.00001395912113	4	Hideki Moronuki
0.00001395912113	4	Ailin Graef
0.00001395912113	4	Larisa Trembovler
0.00001395912113	4	Guiding Light
0.00001395912113	4	Yusuf Kalla
0.00001395912113	4	Dietrich Mateschitz

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001395912113	4	Kadhem Al-Rawi
0.00001395912113	4	All-Stars Starters
0.00001395912113	4	Alix McAlister
0.00001395912113	4	Lhadon Tethong
0.00001395912113	4	Mount Prospect
0.00001395912113	4	Andre Agassi
0.00001395912113	4	Mo Siegel
0.00001395912113	4	serial killer
0.00001395912113	4	Raman Raghav
0.00001395912113	6	Vladislav Listyev
0.00001395912113	4	Skip Spence
0.00001395912113	6	Jack Ma
0.00001395912113	4	Paula Radcliffe
0.00001395912113	4	Christopher Hitchens
0.00001353611697	8	Leona Lewis
0.00001342223186	5	Ahmed Qorei
0.00001342223186	5	Thabo Mbeki
0.00001342223186	5	Qin Shi
0.00001338622224	89	era um
0.00001325986977	32	pode ter
0.00001322443040	6	Melinda Duckett
0.00001322443040	6	Satellite Radio
0.00001316898215	10	Stephen Harper
0.00001307154616	18	partido político
0.00001292511206	10	aquecimento global
0.00001276262446	8	British Airways
0.00001256320866	6	batatas fritas
0.00001256320866	6	Ramzan Kadyrov
0.00001246350075	5	Berni Ahern
0.00001246350075	5	Benjamin Netanyahu
0.00001246350075	5	Quinta Emenda
0.00001246350075	5	Princípios Valdez
0.00001246350075	5	Giles Chichester
0.00001240810707	4	Elvis Presley
0.00001240810707	4	Moqtada al-Sadr

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001240810707	4	Got Talent
0.00001240810707	4	Ban Ki-moon
0.00001240810707	4	Gravetye Manor
0.00001240810707	4	Gertrude Jekyll
0.00001240810707	4	Yves Engler
0.00001240810707	4	Sunshine Coast
0.00001240810707	4	barbárie medieval
0.00001240810707	4	Jacqui Smith
0.00001240810707	4	Great Yarmouth
0.00001226900258	28	têm sido
0.00001215488101	16	of the
0.00001214496478	71	não é
0.00001196496032	6	Alberto Gonzales
0.00001196496032	6	Alberto Tomba
0.00001196496032	6	Big Brother
0.00001179730043	104	e os
0.00001179305036	14	eleição presidencial
0.00001163260094	5	Associated Press
0.00001163260094	5	Marc Ravalomanana
0.00001159680778	18	duas semanas
0.00001152403456	99	em seu
0.00001142109886	6	Wikimedia Foundation
0.00001126035750	11	América Latina
0.00001116729618	4	Heydar Aliyev
0.00001116729618	4	Radovan Karadzic
0.00001116729618	4	Derek Plumbly
0.00001116729618	4	Aptis Comunicações
0.00001116729618	4	Vaclav Havel
0.00001116729618	4	Maurizio Bevilacqua
0.00001116729618	4	Estrelas Foods
0.00001116729618	4	Michelle Steele
0.00001116729618	4	Julia Redd
0.00001116729618	4	Nicole Wong
0.00001116729618	4	coleiras elétricas
0.00001116729618	4	Estelle Getty

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001116729618	4	Jenson Button
0.00001116729618	4	Bowery Poetry
0.00001116729618	4	Lulu Xingwana
0.00001116729618	4	Andrei Kozyrev
0.00001116729618	4	caçadores furtivos
0.00001116729618	4	Idi Amin
0.00001116729618	4	Hilda Solis
0.00001116729618	4	Jade Goody
0.00001116729618	4	Jalal Talabani
0.00001108518427	9	produtos químicos
0.00001103220802	7	mudanças climáticas
0.00001092452931	6	Super Bowl
0.00001090556270	5	Dave McCool
0.00001076430453	93	que se
0.00001056365909	28	pessoas morreram
0.00001046934085	3	Calista Flockhart
0.00001046934085	3	Francis Ricciardone
0.00001046934085	3	Hector Oquelí
0.00001046934085	3	Tel Aviv
0.00001046934085	6	Ted Stevens
0.00001046934085	3	Huiyuan Juice
0.00001046934085	3	Société Générale
0.00001046934085	3	Países Baixos
0.00001046934085	3	Piazza dei
0.00001046934085	3	Lleyton Hewitt
0.00001046934085	3	Niel Tupas
0.00001046934085	3	Burkina Faso
0.00001046934085	3	Vantagem Legal
0.00001046934085	3	Learning Summit
0.00001046934085	3	WealthBridge Connect
0.00001046934085	3	Aer Lingus
0.00001046934085	3	Terry McAuliffe
0.00001042625718	11	James Bond
0.00001033158605	15	Partido Trabalhista
0.00001026405971	5	Jack Ruby

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00001026405971	5	Jacques Delors
0.00001026405971	5	ursos polares
0.00001015208818	4	Sergey Brin
0.00001015208818	4	Land Rover
0.00001015208818	4	Rodrigo Avila
0.00001015208818	4	Giampaolo Giuliani
0.00001015208818	4	Patrick Doohan
0.00001015208818	4	All-Star Game
0.00001015208818	4	Preston Burch
0.00001015208818	4	Lord Carnarvon
0.00001015208818	4	Roger Federer
0.00001015208818	4	Doris Lessing
0.00001015208818	4	Criss Angel
0.00001015208818	4	Charing Cross
0.00001005056674	6	Alan Johnston
0.00000980500135	35	seu filho
0.00000974440991	17	seis meses
0.00000969383382	5	Pierre Beregovoy
0.00000969383382	5	Leona Helmsley
0.00000966400694	6	Neil Armstrong
0.00000966400694	6	telefones celulares
0.00000966400694	6	Ilhas Salomão
0.00000963393177	93	como o
0.00000933131378	86	contra o
0.00000930608076	4	JK Rowling
0.00000930608076	4	Colin Powell
0.00000930608076	4	Circuit Court
0.00000930608076	6	horário nobre
0.00000930608076	4	Kofi Annan
0.00000930608076	4	Russell Dunham
0.00000930608076	4	Estrela Alimentos
0.00000930608076	4	Elmwood Avenue
0.00000930608076	4	Joseph Dault
0.00000930608076	6	Oceano Antártico
0.00000930608076	4	Joba Chamberlain

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000930608076	4	Jason MacIntyre
0.00000930608076	4	Toro Rosso
0.00000930608076	6	Virginia Tech
0.00000930608076	4	Arthur Virgílio
0.00000930608076	4	Sondhi Limthongkul
0.00000930608076	4	Jurgen Schneider
0.00000930608076	4	Joachim Johansson
0.00000930608076	4	Modelo Padrão
0.00000930608076	4	Irene Khan
0.00000930608076	4	Marion Bartoli
0.00000930608076	4	Lycos Europe
0.00000930608076	4	tigres siberianos
0.00000930608076	4	assessoria jurídica
0.00000930608076	4	Salvatore Lo
0.00000930608076	4	Lo Piccolo
0.00000930608076	4	Alexandre Borovik
0.00000924320193	7	moeda estrangeira
0.00000918363185	5	Lana Clarkson
0.00000918363185	5	Brigham Young
0.00000913687927	12	homens armados
0.00000913365056	91	contra a
0.00000909091796	85	que não
0.00000897372047	3	Audrey Seiler
0.00000897372047	3	Tansu Ciller
0.00000897372047	3	Wolfgang von
0.00000897372047	3	Georgi Markov
0.00000897372047	3	Maurice Strong
0.00000897372047	3	Viktor Yanukovych
0.00000897372047	3	Cate Blanchett
0.00000897372047	3	castelo Crathes
0.00000897372047	3	NetIP members
0.00000897372047	6	Boston Celtics
0.00000876919148	14	Partido Liberal
0.00000872445071	5	Boeing 737
0.00000872445071	5	Star Trek

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000865705169	17	já havia
0.00000863379591	50	entre os
0.00000859022839	4	Pete Wilson
0.00000859022839	4	Serra Leoa
0.00000858608018	110	em que
0.00000858009844	77	para as
0.00000847518004	17	duas vezes
0.00000845510476	19	logo após
0.00000837547213	6	Steve Jobs
0.00000810529582	6	ensino médio
0.00000804311276	11	ataques terroristas
0.00000797664052	4	Gerhard Schroeder
0.00000797664052	8	Banco Central
0.00000797664052	4	Mel Sembler
0.00000797664052	4	Ronald Reagan
0.00000797664052	4	Conservação Ambiental
0.00000797664052	4	Ed Stelmach
0.00000797664052	4	Gerald Posner
0.00000797664052	4	Robbie Fowler
0.00000797664052	4	Galvarino Apablaza
0.00000797664052	4	Claire Danes
0.00000793131858	5	instituições financeiras
0.00000793131858	5	Melhor Ator
0.00000789344722	74	diz que
0.00000785200518	3	Sun Microsystems
0.00000785200518	3	Sudeste Asiático
0.00000785200518	3	Pan Am
0.00000785200518	3	Gavin Newsom
0.00000785200518	3	Pai Btiz
0.00000784380154	133	de sua
0.00000759996556	14	nesta quarta-feira
0.00000746569185	73	que eles
0.00000744486442	4	dessas economias
0.00000739012285	6	Kim Beazley
0.00000739012285	6	iPod nano

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000728867826	56	com as
0.00000727656288	7	Greater Manchester
0.00000727037559	5	Margaret Thatcher
0.00000727037559	10	New Hampshire
0.00000720470734	8	crescimento econômico
0.00000715121223	71	que ela
0.00000712496785	7	relações sexuais
0.00000699360908	74	após o
0.00000697956057	2	Zack Urlocker
0.00000697956057	2	Viroj Laohaphan
0.00000697956057	3	Iyad Allawi
0.00000697956057	6	trabalhadores humanitários
0.00000697956057	2	decepcionante 112.000
0.00000697956057	2	Alaattin Çakici
0.00000697956057	4	Football Club
0.00000697956057	2	Josko Damic
0.00000697956057	2	Quentin Tarantino
0.00000697956057	2	Mein Kampf
0.00000697956057	2	Padraig Pearse
0.00000697956057	2	dolce vita
0.00000697956057	2	Brigadas Vermelhas
0.00000697956057	2	Goetz Friedrich
0.00000697956057	2	Kurt Cobain
0.00000697956057	2	Clermont Ferrand
0.00000697956057	2	Yom Kippur
0.00000697956057	2	Rosinha Matheus
0.00000697956057	2	Borislav Shervinsky
0.00000697956057	2	Oppong Manneh
0.00000697956057	2	Malcolm Brabant
0.00000697956057	2	Camille Pissarro
0.00000697956057	3	Gilda Flores
0.00000697956057	2	Jake Peavy
0.00000697956057	2	Les Combes
0.00000697956057	2	Regina Schueller
0.00000697956057	2	Henrik Larsson

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000697956057	2	Zeev Bielsky
0.00000697956057	2	Abdelaziz Bouteflika
0.00000697956057	2	Lindsay Ellis
0.00000697956057	3	Samir Geagea
0.00000697956057	2	tutela síria
0.00000697956057	2	Sendero Luminoso
0.00000697956057	2	Jubiz Hazvumb
0.00000697956057	2	Beatriz Iero
0.00000697956057	2	Industri Kapital
0.00000697956057	2	Barrow Hanley
0.00000697956057	2	Yevgenia Timoshenko
0.00000697956057	2	Helicobacter pylori
0.00000697956057	2	Karl Rove
0.00000697956057	2	Muktar Disse-Ibrahim
0.00000697956057	4	Formula One
0.00000697956057	3	Highway Patrol
0.00000697956057	2	Beth Ditto
0.00000697956057	2	Sara Hiom
0.00000697956057	2	Billy Connolly
0.00000697956057	2	Deacon Brodie
0.00000697956057	2	Nikola Gruevski
0.00000697956057	2	Connie Beauchamp
0.00000697956057	2	Yacht Charters
0.00000697956057	2	Liam Neeson
0.00000697956057	2	Humphrey Bogart
0.00000697956057	2	Ratu Josefa
0.00000697956057	2	Slumdog Millionaire
0.00000697956057	2	Yisrael Beiteinu
0.00000697956057	4	Tesla Motors
0.00000697956057	3	Kenneth Branagh
0.00000697956057	2	Puerto Cabezas
0.00000697956057	2	Baz Luhrman
0.00000697956057	2	alter ego
0.00000697956057	2	Kiryat Arba
0.00000697956057	2	Andriy Shevchenko

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000697956057	2	Midge Ure
0.00000697956057	2	Aurore Paquiss
0.00000697956057	2	Budleigh Salterton
0.00000697956057	2	lona infinito
0.00000697956057	2	Reinventando Comics
0.00000697956057	2	Markus Mller
0.00000697956057	2	Technischen Hochschule
0.00000697956057	2	Irma Goldberg
0.00000697956057	2	Agua Dulce
0.00000697956057	2	Ulan Bator
0.00000697956057	2	Willy Claes
0.00000697956057	2	ex-repúblicas soviéticas
0.00000697956057	2	Plaid Cymru
0.00000697956057	2	RJR Nabisco
0.00000697956057	2	Sento Shosho
0.00000697956057	2	Bulelani Ngcuka
0.00000697956057	2	Forces Nouvelles
0.00000697956057	2	simbolismo religioso
0.00000697956057	2	Pledge Room
0.00000697956057	2	climate change
0.00000697956057	2	entrepreneurial ecosystem
0.00000697956057	3	Wells Fargo
0.00000697956057	2	Attila Ekici
0.00000697956057	2	Leroy Chiao
0.00000697956057	2	Salizhan Sharipov
0.00000697956057	2	Point Comfort
0.00000697956057	2	Horatiu Nastase
0.00000697956057	2	Jeanna Giese
0.00000697956057	2	Nagashi Furukawa
0.00000697956057	2	Kaew Panjapetchkaew
0.00000697956057	2	Prim Palver
0.00000697956057	2	Purnomo Yusgiantoro
0.00000697956057	2	credor hipotecário
0.00000697956057	2	Claudio Ranieri
0.00000697956057	2	Mufi Hannemann

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000697956057	2	Kamal Dahal
0.00000697956057	3	Lago Kivu
0.00000695081781	92	e um
0.00000685642453	35	disse à
0.00000679092363	6	reformas econômicas
0.00000677899834	13	ataque terrorista
0.00000672638134	78	para uma
0.00000670585223	7	Bob Denver
0.00000665111065	9	pediu desculpas
0.00000665032803	90	a uma
0.00000662735420	122	de seu
0.00000661221520	6	lucro líquido
0.00000661221520	6	Bob Rae
0.00000661221520	6	números primos
0.00000657689361	7	floresta tropical
0.00000656899783	4	Mel Gibson
0.00000656899783	4	Raymond Leblanc
0.00000656899783	4	Senador Dunn
0.00000656899783	4	Lori Mitchell
0.00000656899783	4	Flat Top
0.00000656899783	4	Susan Linn
0.00000656899783	4	assistentes sociais
0.00000654542509	14	esta semana
0.00000646255603	5	Philip Morris
0.00000644267129	6	Hyde Park
0.00000642929081	62	por uma
0.00000638131223	24	poderia ser
0.00000635429933	68	depois que
0.00000631366584	16	suas ações
0.00000631277862	27	vai ser
0.00000628160433	3	Emile Lahoud
0.00000628160433	3	shopping center
0.00000628160433	3	Uri Geller
0.00000628160433	3	Baker Children
0.00000628160433	3	Sergey Orlovskiy

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000624098266	13	mil dólares
0.00000623175038	5	Boston Globe
0.00000623175038	5	General Motors
0.00000622234393	63	do presidente
0.00000620405353	4	Tommy Thompson
0.00000620405353	4	Hassan Turabi
0.00000620405353	4	Akhmad Kadyrov
0.00000620405353	4	Vladislav Doronin
0.00000620405353	4	Trenton Duckett
0.00000620405353	4	Carroll Campbell
0.00000620405353	4	Zinedine Zidane
0.00000620405353	4	Augusto Pinochet
0.00000612372696	48	uma nova
0.00000603637636	8	alto perfil
0.00000603637636	8	meu marido
0.00000597591225	73	durante a
0.00000587752447	4	Emerald InTouch
0.00000587752447	8	David Cameron
0.00000587752447	4	Sir Clement
0.00000587752447	4	extremista islâmico
0.00000587752447	4	Morris Iemma
0.00000587752447	4	has been
0.00000587752447	4	envolvendo motociclistas
0.00000571054943	3	monarca jordaniano
0.00000571054943	3	Rolling Stone
0.00000571054943	3	Pedro Quintanar
0.00000571054943	3	party supply
0.00000571054943	3	Il Manifesto
0.00000571054943	3	Glenn Morrison
0.00000570852217	32	as suas
0.00000568026735	54	como uma
0.00000558364809	2	Buzz Aldrin
0.00000558364809	2	Vue Cinemas
0.00000558364809	2	Charlton Heston
0.00000558364809	2	Keith Maupin

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000558364809	2	formato Blu-ray
0.00000558364809	2	Sony BMG
0.00000558364809	2	Rafael Benitez
0.00000558364809	4	Steve Price
0.00000558364809	2	Sally Struthers
0.00000558364809	2	Kimi Raikkonen
0.00000558364809	4	Evo Morales
0.00000558364809	2	Shin Tae-seop
0.00000558364809	2	toques polifônicos
0.00000558364809	2	Alexandra Holzer
0.00000558364809	2	baby boom
0.00000558364809	2	Adil Kalbani
0.00000558364809	2	Tsuyoshi Kusanagi
0.00000558364809	2	ski run
0.00000558364809	2	Kathleen Sebelius
0.00000558364809	2	união monetária
0.00000558364809	2	Jihad Islâmica
0.00000558364809	2	economistas previam
0.00000558364809	2	Maggie Dempster
0.00000558364809	2	Ibrahim Sofu
0.00000558364809	2	Montanhas Azuis
0.00000558364809	2	Symbian OS
0.00000558364809	2	HOT 97
0.00000558364809	2	towards children
0.00000558364809	2	Laurie Garner
0.00000558364809	2	Toni Hoffman
0.00000558364809	2	Isaac Asimov
0.00000558364809	4	Jeopardy !
0.00000558364809	2	Bosco Ntaganda
0.00000555109546	72	a ser
0.00000545160265	17	havia sido
0.00000531776050	8	Eu estou
0.00000531776050	4	Murray Hill
0.00000531776050	4	Marina Souza
0.00000531776050	4	Barnaby Joyce

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000525519818	8	bomba explodiu
0.00000525437417	60	dizem que
0.00000523467043	3	Sonia Gandhi
0.00000523467043	3	Grand Prix
0.00000523467043	3	Casey Stoner
0.00000523467043	3	sales plan
0.00000523155995	29	não vai
0.00000517084209	41	sobre os
0.00000513202986	5	parque temático
0.00000507604409	4	gripe aviária
0.00000507604409	4	Ian Fleming
0.00000507604409	4	lhe dissera
0.00000507604409	4	Naomi Campbell
0.00000507604409	4	General Store
0.00000495162294	33	não está
0.00000492674826	6	militares indonésios
0.00000486937506	63	e as
0.00000485534656	4	Fundo Global
0.00000485534656	4	motor esquerdo
0.00000485534656	4	Steven Spielberg
0.00000485534656	4	Thunder Bay
0.00000485534656	4	vulcão Ubinas
0.00000485534656	4	Wendy Portillo
0.00000485534656	4	Tata Steel
0.00000485534656	4	Jim Callahan
0.00000485534656	4	Mark Cuban
0.00000485104192	28	as mulheres
0.00000483200347	3	Alpes italianos
0.00000483200347	3	Timor Leste
0.00000483200347	3	criminalidade violenta
0.00000483200347	3	Católica Romana
0.00000481478401	61	que está
0.00000478387165	29	muito mais
0.00000474021226	72	para um
0.00000465304038	2	Klaus Wowereit

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000465304038	2	Mani Tripathi
0.00000465304038	2	Madhu Mani
0.00000465304038	2	Olive Madison
0.00000465304038	2	Front Row
0.00000465304038	2	Ben Kinsella
0.00000465304038	2	Will Carling
0.00000465304038	2	Zhirun Yuan
0.00000465304038	2	Sherman Hartley
0.00000465304038	2	pole position
0.00000465304038	2	Kirk Kerkorian
0.00000465304038	2	Silicon Graphics
0.00000465304038	2	Nicol Williamson
0.00000465304038	2	Salvo Lima
0.00000465304038	2	Ben Stiller
0.00000465304038	4	Mahmoud Abbas
0.00000465304038	2	Achille Occhetto
0.00000465304038	2	Ursa Maior
0.00000465304038	2	anglo-holandesa Corus
0.00000465304038	2	Sergei Sidorsky
0.00000465304038	2	Donald Tusk
0.00000465304038	2	Ursos pardos
0.00000465304038	2	toxic chemicals
0.00000465304038	2	Toby Harnden
0.00000465304038	2	Ezer Weizman
0.00000465304038	2	Toby Gascon
0.00000465304038	2	vaso sanitário
0.00000465304038	2	Pablo Neruda
0.00000465304038	2	Gossip Girl
0.00000459181592	5	gravemente ferido
0.00000456973794	16	eles estavam
0.00000432760453	25	seu primeiro
0.00000425871485	6	Igreja Batista
0.00000423902111	23	três anos
0.00000409809036	8	pelas forças
0.00000405788387	5	pedir desculpas

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000398832026	2	Todd Chaves
0.00000398832026	2	Intel Pentium
0.00000398832026	2	delegação marroquina
0.00000398832026	2	Perez Quiron
0.00000398832026	2	Madeleine Albright
0.00000398832026	2	Marilyn Monroe
0.00000398832026	2	Alexander Downer
0.00000398832026	2	Raul Reyes
0.00000398832026	2	Barry Rogerson
0.00000398832026	2	considerada insegura
0.00000398832026	2	Lester Piggott
0.00000398832026	2	Natasha Richardson
0.00000398832026	2	Rita Levi-Montalcini
0.00000398832026	2	years ago
0.00000398832026	2	esclerose múltipla
0.00000398832026	2	Virgin Atlantic
0.00000398832026	2	galáxia anã
0.00000398832026	2	Max Purnell
0.00000398832026	2	Marilyn Manson
0.00000398832026	2	Metin Kaplan
0.00000398832026	2	Mohamed Fulayfel
0.00000398832026	2	míssil balístico
0.00000398832026	2	Labrador Inuit
0.00000398832026	2	Bairro Francês
0.00000398832026	2	Mona Lisa
0.00000398832026	2	Nadya Suleman
0.00000396565929	25	as forças
0.00000395302550	8	F 1
0.00000394139897	12	nas eleições
0.00000392600259	6	estamos aqui
0.00000392049151	34	os seus
0.00000379628523	51	disseram que
0.00000379323933	5	Apple Computer
0.00000358948819	6	F- 16
0.00000356256919	66	que um

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000356021337	49	para se
0.00000349998845	60	e uma
0.00000348978028	2	meteorito marciano
0.00000348978028	2	Walter Cronkite
0.00000348978028	2	Korkmaz Yigit
0.00000348978028	2	London Underground
0.00000348978028	2	Nicole Kidman
0.00000348978028	2	concedido anualmente
0.00000348978028	2	jatos corporativos
0.00000348978028	2	Gerald Schatten
0.00000348978028	2	profeta Maomé
0.00000348978028	2	Old Mutual
0.00000348978028	2	Ronnie Gilbert
0.00000348978028	2	Fernando Alonso
0.00000348978028	2	119 universidades
0.00000348978028	2	Kia Pride
0.00000348978028	2	alimento básico
0.00000348978028	2	Tyson Gay
0.00000348978028	2	Brendan Fraser
0.00000348978028	2	prisões secretas
0.00000348978028	2	Salvatore Gravano
0.00000348978028	2	Andrei Kokoshin
0.00000348978028	2	cabo Bravo
0.00000348978028	2	Granada Teatro
0.00000348978028	3	will be
0.00000345452986	49	e outros
0.00000332740296	25	entre as
0.00000330610760	3	Albert Reynolds
0.00000330610760	3	Privacy International
0.00000330610760	3	Josh Schwartz
0.00000323844597	43	um novo
0.00000319065612	4	Al Jazeera
0.00000316562068	50	e não
0.00000314080216	3	companhias aéreas
0.00000312858106	24	não tem

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000310202677	2	Wildlife Conservation
0.00000310202677	2	Hun Sen
0.00000310202677	2	ciclistas tirou
0.00000310202677	2	Jennifer Watts
0.00000310202677	2	Danny Boyle
0.00000310202677	2	Hudson Valley
0.00000310202677	2	ambientalmente saudáveis
0.00000310202677	2	Small Heath
0.00000310202677	2	bacia amazônica
0.00000310202677	2	Universal Pictures
0.00000310202677	2	transtorno bipolar
0.00000310202677	2	wood chips
0.00000303872025	8	Partido Socialista
0.00000297574161	24	como eles
0.00000296740404	34	não foi
0.00000291422134	20	sua vida
0.00000285527472	3	agressão sexual
0.00000279182404	2	Escritor Samir
0.00000279182404	2	Mauricio Pineda
0.00000279182404	2	Marina Petrella
0.00000279182404	2	dieta mediterrânea
0.00000279182404	2	Michel Rocard
0.00000279182404	2	Amanda Mealing
0.00000279182404	2	Carrie Prejean
0.00000279182404	2	Muammar Gaddafi
0.00000279182404	2	Katherine Heigl
0.00000279182404	2	raios X
0.00000279182404	2	JP Losman
0.00000279182404	2	SBC Communications
0.00000279182404	2	Soto Toro
0.00000279182404	2	Alfred Hitchcock
0.00000279182404	2	Suresh Joachim
0.00000279182404	2	Rebecca MacKinnon
0.00000279182404	2	substâncias dopantes
0.00000279182404	2	Cristiano Ronaldo

Table B.2: (continued)

Mutual Expectation	Frequency	**MWU**
0.00000277921527	46	que foram
0.00000274435342	57	e da
0.00000271226440	15	informações sobre
0.00000267935729	9	cinco vezes
0.00000256695080	11	muito tempo
0.00000253802204	2	Post Office
0.00000253802204	2	Van Meeteren
0.00000253802204	2	Ingrid Bergman
0.00000253802204	2	acessórios digitais
0.00000253802204	2	Jorge Lorenzo
0.00000253802204	2	Amanda Knox
0.00000253802204	2	Shenzhou V
0.00000253802204	2	Christiaan Huygens
0.00000247824960	28	no dia
0.00000245331285	31	ele é
0.00000241600173	3	Fidel Castro
0.00000232652019	3	sistemas operacionais
0.00000224565179	19	ele tem
0.00000216607054	3	comissão parlamentar
0.00000213128124	39	que vai
0.00000210703706	4	novas tecnologias
0.00000208769143	13	algumas das
0.00000202632395	3	Financial Times
0.00000201011335	6	in the
0.00000197524469	33	trabalha para
0.00000195919188	45	que uma
0.00000195917482	4	partidos políticos
0.00000194307722	24	contra os
0.00000193399137	15	também são
0.00000179474409	3	setor privado
0.00000172361092	27	tem uma
0.00000170059809	69	e de
0.00000156255408	37	começou a

Appendix C

Sample form submitted to the *"Turkers"* in CrowdFlower

Building a corpus of pairs Text -> Hypothesis, to learn Recognizing Textual Entailment by Generality

Instructions

When people talk and write they convey both explicit and implicit information. People are in general capable to infere the implicit part. For example, when one tells that they have stained their white shirt, we can deduce that the white one is not the only one that the speaker has.

We are interested in learning to recognise, given a pair of sentences - the Text T and the Hypothesis H, whether H can be infered provided the T is true. This task is called Recognition of Textual Entailment and is a key task for many natural language processing (NLP) applications.

Thus, Recognizing Textual Entailment consists in determining whether an entailment relation holds between a text T and a hypothesis H. It is said that entailment relation holds between T and H when the meaning of H can be inferred from T.

Informally, T entails H if, typically, a human reading T would infer that H is most likely true based on the truth of T.

In logic deductive reasoning is a kind of inference in which the conclusion is of no greater generality than the premises. This is why we put all cases of TE in two categories - TE but not Generality which is equivalent to deductive reasoning and TE by generality, where the Hypothesis H is in some sense more general than the Text T. For example, from T below one can conclude that Unterweger was not killed as no one has access to him inside the cell, so it was a suicide, but H does not elaborate on the exat means in contrast to T. Thus, the realtion between T - H is TE by Generality.

T: Unterweger was found hanging from his cell roof less than 24 hours after being sentenced to life for a second time.

H: Unterweger committed suicide.

On the other side the following pair T - H is TE but not Generality as H is true given T, but H is no more general with respect to the details it has in common with T.

T: With Linnaeus as professor, a period in Uppsala began where nature science was much esteemed.

H: Linnaeus was a professor in Uppsala.

All pairs T - H in this job come from RTE1 through RTE5 PASCAL Challenges.

T: An avalanche has struck a popular skiing resort in Austria, killing at least 11 people.

H: Humans died in an avalanche.

Choose one

- ◯ Entailment by Generality
- ◯ Entailment, but not Generality
- ◯ Other

T: Tea also contains anti-oxidants that help to reduce the chances of heart disease and cancer.

H: Tea protects from some diseases.

Choose one

- ◯ Entailment by Generality
- ◯ Entailment, but not Generality
- ◯ Other

T: Nelson Mandela's Long Walk to Freedom began as scraps of paper, buried under the floor of his prison cell.

H: Nelson Mandela's autobiography is called "The Long Walk to Freedom".

Choose one

- ◯ Entailment by Generality
- ◯ Entailment, but not Generality
- ◯ Other

Appendix D

Sample Web Frequencies for Calculations

Pair T-H extracted from RTE-3 test set.

<pair id="217" entailment="YES" task="IR" length="short">
<t>Pierre Beregovoy, apparently left no message when he shot himself with a borrowed gun.</t>
<h>Pierre Beregovoy commits suicide.</h>
</pair>

D.1 All Words

Table D.1: Web frequencies for calculations with ***All Words***

T	**Frequency**	*H*	**Frequency**
"Pierre"	61500000	*"Pierre"*	61500000
"Beregovoy"	16900	*"Beregovoy"*	16900
"apparently"	29600000	*"commits"*	401000000
"left"	249000000	*"suicide"*	1880000000
"no"	1710000000		
"message"	333000000		
"when"	802000000		
"he"	402000000		
"shot"	90200000		
"himself"	46500000		
"with"	1840000000		
"a"	3750000000		
"borrowed"	4710000		
"gun"	59200000		

Table D.2: Web frequencies for calculations with ***All Words***

		Frequency	
T	**H**	**T ∩ H**	**H ∩ T**
"Pierre"	*"Beregovoy"*	11500	11600
"Pierre"	*"commits"*	59200	60100
"Pierre"	*"suicide"*	1560000	1670000
"Beregovoy"	*"commits"*	32	33
"Beregovoy"	*"suicide"*	1220	1220
"apparently"	*"Pierre"*	1460000	1430000
"apparently"	*"Beregovoy"*	231	230
"apparently"	*"commits"*	3320000	3040000
"apparently"	*"suicide"*	17300000	16700000
"left"	*"Pierre"*	30900000	22300000
"left"	*"Beregovoy"*	761	765
"left"	*"commits"*	21400000	21100000
"left"	*"suicide"*	136000000	114000000
"no"	*"Pierre"*	75800000	76700000
"no"	*"Beregovoy"*	4540	4580
"no"	*"commits"*	37700000	39700000
"no"	*"suicide"*	263000000	273000000
"message"	*"Pierre"*	15700000	14400000
"message"	*"Beregovoy"*	1890	1880
"message"	*"commits"*	5700000	1970000
"message"	*"suicide"*	62800000	62800000
"when"	*"Pierre"*	83000000	79600000
"when"	*"Beregovoy"*	1490	1490
"when"	*"commits"*	37800000	39900000
"when"	*"suicide"*	268000000	172000000
"he"	*"Pierre"*	57700000	55900000
"he"	*"Beregovoy"*	1110	1100
"he"	*"commits"*	34100000	35000000
"he"	*"suicide"*	239000000	136000000
"shot"	*"Pierre"*	18800000	15800000
"shot"	*"Beregovoy"*	271	275
"shot"	*"commits"*	22200000	19600000
"shot"	*"suicide"*	108000000	114000000

Table D.2: (continued)

		Frequency	
T	**H**	**T ∩ H**	**H ∩ T**
"himself"	*"Pierre"*	1960000	2000000
"himself"	*"Beregovoy"*	328	329
"himself"	*"commits"*	14900000	7340000
"himself"	*"suicide"*	58000000	46000000
"with"	*"Pierre"*	101000000	97300000
"with"	*"Beregovoy"*	2900	2920
"with"	*"commits"*	43800000	50300000
"with"	*"suicide"*	201000000	293000000
"a"	*"Pierre"*	105000000	98200000
"a"	*"Beregovoy"*	14500	14700
"a"	*"commits"*	44300000	50900000
"a"	*"suicide"*	311000000	294000000
"borrowed"	*"Pierre"*	119000	144000
"borrowed"	*"Beregovoy"*	44	44
"borrowed"	*"commits"*	40800	40900
"borrowed"	*"suicide"*	843000	717000
"gun"	*"Pierre"*	1910000	4360000
"gun"	*"Beregovoy"*	193	193
"gun"	*"commits"*	14000000	11500000
"gun"	*"suicide"*	71000000	72600000

D.2 Without Stop Words

Table D.3: Web frequencies for calculations without ***Stop Words***

T	**Frequency**	*H*	**Frequency**
"Pierre"	61500000	*"Pierre"*	61500000
"Beregovoy"	16900	*"Beregovoy"*	16900
"apparently"	29600000	*"commits"*	401000000
"left"	249000000	*"suicide"*	1880000000
"message"	333000000		
"when"	802000000		
"shot"	90200000		
"himself"	46500000		
"borrowed"	4710000		
"gun"	59200000		

Table D.4: Web frequencies for calculations without ***Stop Words***

		Frequency	
T	**H**	**T ∩ H**	**H ∩ T**
"Pierre"	*"Beregovoy"*	11500	11600
"Pierre"	*"commits"*	59200	60100
"Pierre"	*"suicide"*	1560000	1670000
"Beregovoy"	*"commits"*	32	33
"Beregovoy"	*"suicide"*	1220	1220
"apparently"	*"Pierre"*	1460000	1430000
"apparently"	*"Beregovoy"*	231	230
"apparently"	*"commits"*	3320000	3040000
"apparently"	*"suicide"*	17300000	16700000
"left"	*"Pierre"*	30900000	22300000
"left"	*"Beregovoy"*	761	765
"left"	*"commits"*	21400000	21100000
"left"	*"suicide"*	136000000	114000000
"message"	*"Pierre"*	15700000	14400000
"message"	*"Beregovoy"*	1890	1880
"message"	*"commits"*	5700000	1970000
"message"	*"suicide"*	62800000	62800000
"when"	*"Pierre"*	83000000	79600000

Table D.4: (continued)

		Frequency	
T	**H**	**T ∩ H**	**H ∩ T**
"when"	*"Beregovoy"*	1490	1490
"when"	*"commits"*	37800000	39900000
"when"	*"suicide"*	268000000	172000000
"shot"	*"Pierre"*	18800000	15800000
"shot"	*"Beregovoy"*	271	275
"shot"	*"commits"*	22200000	19600000
"shot"	*"suicide"*	108000000	114000000
"himself"	*"Pierre"*	1960000	2000000
"himself"	*"Beregovoy"*	328	329
"himself"	*"commits"*	14900000	7340000
"himself"	*"suicide"*	58000000	46000000
"borrowed"	*"Pierre"*	119000	144000
"borrowed"	*"Beregovoy"*	44	44
"borrowed"	*"commits"*	40800	40900
"borrowed"	*"suicide"*	843000	717000
"gun"	*"Pierre"*	1910000	4360000
"gun"	*"Beregovoy"*	193	193
"gun"	*"commits"*	14000000	11500000
"gun"	*"suicide"*	71000000	72600000

D.3 With MultiWords Units

Table D.5: Web frequencies for calculations with ***MultiWords Units***

T	**Frequency**	*H*	**Frequency**
"Pierre Beregovoy"	8300	*"Pierre Beregovoy"*	8300
"apparently"	29600000	*"commits"*	401000000
"left"	249000000	*"suicide"*	1880000000
"no"	1710000000		
"message"	333000000		
"when"	802000000		
"he"	402000000		
"shot"	90200000		
"himself"	46500000		
"with a"	519000000		
"borrowed"	4710000		
"gun"	59200000		

Table D.6: Web frequencies for calculations with ***MultiWords Units***

		Frequency	
T	**H**	**T ∩ H**	**H ∩ T**
"Pierre Beregovoy"	*"commits"*	60	58
"Pierre Beregovoy"	*"suicide"*	2110	2090
"apparently"	*"Pierre Beregovoy"*	144	143
"apparently"	*"commits"*	3320000	3040000
"apparently"	*"suicide"*	17300000	16700000
"left"	*"Pierre Beregovoy"*	1190	1190
"left"	*"commits"*	21400000	21100000
"left"	*"suicide"*	136000000	114000000
"no"	*"Pierre Beregovoy"*	5850	5830
"no"	*"commits"*	37700000	39700000
"no"	*"suicide"*	263000000	273000000
"message"	*"Pierre Beregovoy"*	3120	3120
"message"	*"commits"*	5700000	1970000
"message"'	*"suicide"*	62800000	62800000
"when"	*"Pierre Beregovoy"*	1710	1710

Table D.6: (continued)

		Frequency	
T	**H**	**T ∩ H**	**H ∩ T**
"when"	*"commits"*	37800000	39900000
"when"	*"suicide"*	268000000	283000000
"he"	*"Pierre Beregovoy"*	1520	1520
"he"	*"commits"*	34100000	35000000
"he"	*"suicide"*	239000000	136000000
"shot"	*"Pierre Beregovoy"*	307	308
"shot"	*"commits"*	22200000	19600000
"shot"	*"suicide"*	108000000	114000000
"himself"	*"Pierre Beregovoy"*	514	517
"himself"	*"commits"*	14900000	7340000
"himself"	*"suicide"*	58000000	46000000
"with a"	*"Pierre Beregovoy"*	1160	1150
"with a"	*"commits"*	23400000	23300000
"with a"	*"suicide"*	95900000	89600000
"borrowed"	*"Pierre Beregovoy"*	37	37
"borrowed"	*"commits"*	40800	40900
"borrowed"	*"suicide"*	843000	717000
"gun"	*"Pierre Beregovoy"*	220	220
"gun"	*"commits"*	14000000	11500000
"gun"	*"suicide"*	71000000	72600000

Index

AAM - Asymmetric Association Measures, vii, 14, 55, 56, 64, 67
AC - Accuracy, 33, 39, 41, 56, 63, 68, 100
AIS - Asymmetric InfoSimba Similarity, 13, 14, 43, 58, 101
AISs - Simplified Asymmetric InfoSimba Similarity, 14
Artificial Intelligence, 1

Bayesian Statistics, 1

Cluster, 7, 19, 23, 101
CrowdFlower, viii, 14, 45–48, 50, 100
crowdsourcing, viii, 45–47

Entailment, 8
Ephemeral Clustering, 7, 101

HITs - Human Intelligence Tasks, 46, 47

IAM - Informative Asymmetric Measure, vii, 13, 14, 43

Lexical, 35
Logic Inference, 35, 36
Logical Implication, 8

ML - Machine Learning, 1, 38, 41
ML - Machine Translation, 2–4, 7, 10, 13, 18, 31, 34, 54
MWU - Multiword Units, 14, 62, 63

Natural Language Processing, 1

P - Precision, 68, 100

RAIS- Recursive Asymmetric InfoSimba Similarity, 101

Semantic, 35
Semantic Kernel Function, 36
SENTA, 14, 63
similarity, 34, 36–41, 51–54
Snippets, 35
Stop Words, 14, 62, 63, 65, 67
Syntactic, 35, 36

Turkers, 45–47, 49, 50

Mesures de Similarité Distributionnelle Asymétrique pour la Détection de l'Implication Textuelle par Généralité

Résumé: Textual Entailment vise à capturer les principaux besoins d'inférence sémantique dans les applications de Traitement du Langage Naturel. Depuis 2005, dans la Textual Entailment reconnaissance tâche (RTE), les systèmes sont appelés à juger automatiquement si le sens d'une portion de texte, le texte - T, implique le sens d'un autre texte, l'hypothèse - H. Cette thèse nous nous intéressons au cas particulier de l'implication, l'implication de généralité. Pour nous, il ya différents types d'implication, nous introduisons le paradigme de l'implication textuelle en généralité, qui peut être définie comme l'implication d'une peine spécifique pour une phrase plus générale, dans ce contexte, le texte T implication Hypothèse H, car H est plus générale que T. Nous proposons des méthodes sans surveillance indépendante de la langue de reconnaissance de l'implication textuelle par la généralité, pour cela, nous présentons une mesure asymétrique informatif appelée Asymmetric simplifié InfoSimba, que nous combinons avec différentes mesures d'association asymétriques à reconnaître le cas spécifique de l'implication textuelle par la généralité.Cette thèse, nous introduisons un nouveau concept d'implication, les implications de généralité, en conséquence, le nouveau concept d'implications de la reconnaissance par la généralité, une nouvelle orientation de la recherche en Traitement du Langage Naturel.

Mots clés: Traitement du Langage Naturel, Implication Textuelle, Reconnaissant l'Implication Textuelle en Généralité, Similarité de Mots, Mesure Asymétrique Informatif, Asymétrique Mesure Association

---

Asymmetric Distributional Similarity Measures to Recognize Textual Entailment by Generality

Abstract: Textual Entailment aims at capturing major semantic inference needs across applications in Natural Language Processing. Since 2005, in the Textual Entailment recognition (RTE) task, systems are asked to automatically judge whether the meaning of a portion of text, the Text - T, entails the meaning of another text, the Hypothesis - H. This thesis we focus a particular case of entailment, entailment by generality. For us, there are various types of implication, we introduce the paradigm of Textual Entailment by Generality, which can be defined as the entailment from a specific sentence towards a more general sentence, in this context, the Text T entailment Hypothesis H, because H is more general than T. We propose methods unsupervised language-independent for Recognizing Textual Entailment by Generality, for this we present an Informative Asymmetric Measure called the Simplified Asymmetric InfoSimba, which we combine with different asymmetric association measures to recognizing the specific case of Textual Entailment by Generality.This thesis, we introduce the new concept of implication, implications by generality, in consequence, the new concept of recognition implications by generality, a new direction of research in Natural Language Processing.

Keywords: Natural Language Processing, Textual Entailment, Recognizing Textual Entailment by Generality, Word Similarity, Informative Asymmetric Measure, Asymmetric Association Measure

MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY