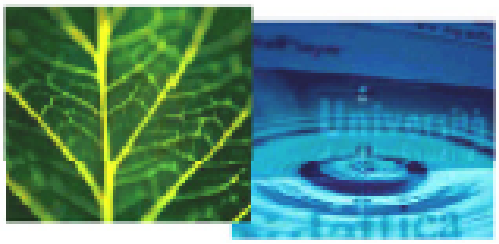
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Component-Based Textual Entailment: a Modular and Linguistically-Motivated Framework for Semantic Inferences

Elena Cabrio

Advisor:

Prof. Bernardo Magnini

Fondazione Bruno Kessler, Human Language Technology Research Unit.

April 2011

Abstract

Textual Entailment (TE) aims at capturing major semantic inference needs across applications in Natural Language Processing. Since 2005, in the TE recognition (RTE) task, systems are asked to automatically judge whether the meaning of a portion of text, the Text, entails the meaning of another text, the Hypothesis. Although several approaches have been experimented, and improvements in TE technologies have been shown in RTE evaluation campaigns, a renewed interest is rising in the research community towards a deeper and better understanding of the core phenomena involved in textual inference. In line with this direction, we are convinced that crucial progress may derive from a focus on decomposing the complexity of the TE task into basic phenomena and on their combination. Analysing TE in the light of the notions provided in logic to define an argument, and to evaluate its validity, the aim of our work is to understand how the common intuition of decomposing TE would allow a better comprehension of the problem from both a linguistic and a computational viewpoint. We propose a framework for component-based TE, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. Five dimensions of the problem are investigated: i) the definition of a component-based TE architecture; ii) the implementation of TE-components able to address specific inference types; iii) the linguistic analysis of the phenomena relevant to component-based TE; iv) the automatic acquisition of knowledge to support component-based entail-

ment judgements; v) the development of evaluation methodologies to assess component-based TE systems capabilities to address single phenomena in a pair.

Keywords

[Natural Language Processing, Semantic Inference, Textual Entailment, Meaning Compositionality]

Acknowledgements

I really wish to thank all the people who supported me, in different ways, during the Ph.D., and who helped me write my dissertation successfully.
I am heartily thankful to my advisor, Bernardo Magnini, who first brought my attention to the topic of semantic inference, and whose supervision encouraged me to develop an understanding of the subject. His support, the constant availability to discuss my thoughts, and the guidance he showed me throughout my dissertation writing have been of great value to me.
I am also indebted to the members of my dissertation committee - Rodolfo Delmonte, Sebastian Padò, Piek Vossen, and Frédérique Segond - for their encouragement and helpful feedback. In particular, I owe special thanks to Frédérique, for giving me the opportunity to spend a period in her group at XRCE in Grenoble, which represents an enriching experience for my professional growth.
I would like to thank Morena Danieli, for instilling in me a love for Computational Linguistics, and for having encouraged me to pursue a Ph.D.
I am grateful to the colleagues of the HLT group at FBK, who have provided a pleasant and stimulating environment to pursue my studies. Particularly, I would like to acknowledge the group of colleagues working on Textual Entailment - Milen Kouylekov, Matteo Negri, and Yashar Mehdad - for the precious collaboration and constructive discussions.
I owe special thanks to Sara Tonelli, Luisa Perenthaler, Bonaventura Coppola, and many others, for the stimulating conversations, the amazing time

spent together in Trento, and their warm companionship.
Finally, I offer my deepest thanks to my family, my parents and my sister Erica, whose love and support has sustained me during these years away from home. And to Daniel, the farthest, the closest.

Contents

1 Introduction **1**

1.1 The Context 1

1.2 The Problem 2

1.3 The Solution 3

1.4 Innovative Aspects/ Contributions 3

1.5 Structure of the Thesis 6

2 Semantic Inferences **9**

2.1 Introduction 9

2.2 Logical argument 10

2.3 Argument evaluation 14

2.3.1 Criterion 1: Truth of premises 14

2.3.2 Criterion 2: Validity and inductive probability . . . 15

2.3.3 Criterion 3: Relevance 18

2.3.4 Criterion 4: The requirement of total evidence . . . 19

2.4 Inductive reasoning 20

2.4.1 Statistical syllogism 21

2.4.2 Statistical generalization 23

2.4.3 Inductive generalization and simple induction . . . 24

2.4.4 Induction by analogy 25

2.4.5 Causality 26

2.5 The notion of Entailment 29

2.6 Computational approaches 30
2.7 Language variability 32
2.8 Textual Entailment 34
2.8.1 Probabilistic Textual Entailment 36
2.8.2 TE and background knowledge 37
2.8.3 Applying argument evaluation criteria to TE pairs . 38
2.9 Conclusion 43

3 RTE **45**
3.1 Introduction 45
3.2 The RTE Evaluation Campaign 46
3.2.1 RTE data sets 50
3.2.2 RTE Approaches 52
3.2.3 Knowledge resources 54
3.2.4 Tools for RTE data preprocessing 55
3.3 Phenomena relevant to inference 56
3.3.1 Sammons *et al.* 2010 [83] 60
3.4 Conclusions 60

4 Framework **63**
4.1 Introduction 63
4.2 Decomposing the TE task 65
4.2.1 Towards total evidence: atomic arguments 65
4.2.2 Linguistic phenomena relevant to inference 71
4.2.3 Entailment rules 73
4.2.4 Contradiction rules 74
4.2.5 Atomic RTE pairs 76
4.3 Dependencies 79
4.4 Architecture 81
4.4.1 TE-components expected behaviour 84

4.4.2 Transformation-based framework 87
4.5 NL for TE-components definition 88
4.5.1 Extended model of NL 89
4.5.2 Defining TE-components using NL relations 92
4.5.3 Entailment Rules and Atomic Edits 93
4.6 TE-components combination 95
4.6.1 Combination based on Natural Logic 95
4.6.2 Order of composition 96
4.6.3 Experimenting NL on RTE pairs 97
4.7 Conclusion 100

5 TE-components implementation 103
5.1 Introduction 103
5.2 EDITS 105
5.3 TE-components 107
5.4 Testing 109
5.4.1 Implemented TE-components 109
5.4.2 Results and error analysis 116
5.5 Combination 120
5.5.1 Compositional strategies 121
5.5.2 Experiments 123
5.6 RTE 124
5.7 Conclusions 125

6 Specialized Data Sets 127
6.1 Introduction 127
6.2 Methodology 129
6.3 Procedure application 131
6.3.1 Entailment pairs 131
6.3.2 Contradiction pairs 133

6.3.3 Unknown pairs 135
6.4 Feasibility Study 136
6.4.1 Inter-annotator agreement 137
6.4.2 Results of the feasibility study 138
6.5 Data Sets creation 141
6.6 Conclusions 144

7 Rules Acquisition **147**
7.1 Introduction 147
7.2 Related Work 150
7.3 General Methodology 152
7.4 Entailment Rules Acquisition 154
7.4.1 Step 1: Preprocessing Wikipedia dumps 154
7.4.2 Step 2: Extraction of entailment pairs 155
7.4.3 Step 3: Extraction of entailment rules 157
7.4.4 Step 4: Rules expansion with minimal context . . . 160
7.5 Experiments and results 162
7.5.1 Evaluation 164
7.5.2 Error Analysis 165
7.6 Conclusion and future work 166

8 Component-based evaluation **169**
8.1 Introduction 169
8.2 Methodology 171
8.2.1 General Method 171
8.2.2 Component Correlation Index (CCI) 173
8.2.3 Component Deviation Index (CDI) 174
8.3 Experiments 176
8.3.1 Data set 176
8.3.2 TE systems 176

8.3.3 Results 178
8.4 contradiction 184
8.5 Conclusion 186

9 Conclusion **189**

Bibliography **195**

A List of Published Papers **209**

List of Tables

4.1 Set $\mathfrak{B}$ of basic semantic relations (MacCartney and Manning 2009 [56]) 90

4.2 Join table for relations in $\mathfrak{B}$ (MacCartney and Manning 2009 [56]) 90

4.3 Application of the NL composition methodology to an *entailment* pair. 98

4.4 Application of the NL composition methodology to a *contradiction* pair. 99

4.5 Application of the NL composition methodology to a *unknown* pair. 100

5.1 Evaluation of the TE-components with respect to neutral behaviour. 116

5.2 Evaluation of the TE-components with respect to positive behaviour. 117

5.3 Evaluation of the TE-components with respect to negative behaviour. 118

5.4 Results comparison over RTE5 data set. 124

5.5 Results on RTE-4 data set 125

6.1 Application of the decomposition methodology to an *entailment* pair. 132

6.2 Application of the decomposition methodology to a *contradiction* pair. 134
6.3 Application of the methodology to an *unknown* pair. . . . 135
6.4 Agreement measures per entailment type 138
6.5 Distribution of phenomena in T-H pairs. 141
6.6 Distribution of the atomic pairs with respect to original E/C/U pairs 142

7.1 Statistics on Wikipedia dumps. 155
7.2 Statistics on pairs similarity. 155
7.3 Statistics on the data sets of entailment rules. 164
7.4 Results of the evaluation of the sets of rules. 164

8.1 Systems' accuracy on phenomena 180
8.2 Evaluation on RTE pairs and on atomic pairs 181
8.3 Evaluation on categories of phenomena 182
8.4 Evaluation on entailment and contradiction pairs 183
8.5 Cooccurrencies of phenomena in contradiction pairs 187

List of Figures

2.1 Venn diagram of the entailment relation 29

3.1 Systems participating in previous RTE challenges (main task) 48
3.2 Systems' performances for the two-way judgement task . . 62
3.3 Baseline for the two-way judgement task 62
3.4 Systems' performances for the three-way judgement task . 62
3.5 RTE data sets with respect to the distribution of logical arguments 62

4.1 Arguments inferential structures 79
4.2 Inferential structure of Example 4.3 80
4.3 Compositional models of atomic arguments 81
4.4 Component-based architecture 88

5.1 EDITS architecture and work-flow 107

7.1 Minimal context LHS rule 168
7.2 Minimal context RHS rule 168

Chapter 1

Introduction

In this Chapter we introduce the context and the motivations underlying the present research work, and provide its positioning in the framework of the research in Natural Language Processing.

1.1 The Context

Textual Entailment (TE) has been proposed as a unifying generic framework for modelling language variability and capturing major semantic inference needs across applications in Natural Language Processing (NLP). Since 2005, in the TE recognition (RTE) task (Dagan *et al.* 2009 [27]), systems are asked to automatically judge whether the meaning of a portion of text, referred as Text (T), entails the meaning of another text, referred as Hypothesis (H). For instance, given the following T-H pairs:

(1.1) T: *Euro-Scandinavian media cheer Denmark vs Sweden draw.*
H: *Denmark and Sweden tie.*

(1.2) T: *Oracle had fought to keep the forms from being released.*
H: *Oracle released a confidential document.*

an RTE system should assign *yes* as the entailment judgement for Example 1.1 (i.e. the meaning of H can be logically derived from the meaning of T),

while it should assign *no* to Example 1.2.

This evaluation provides useful cues for researchers and developers aiming at the integration of TE components in larger applications (see, for instance, the use of a TE engine in the QALL-ME project system[1], the use in relation extraction (Romano *et al.* 2006 [81]), and in reading comprehension systems (Nielsen *et al.* 2009 [74]).

1.2 The Problem

Textual Entailment comes at various levels of complexity and involves almost all linguistic phenomena of natural languages, including lexical, syntactic and semantic variations. Although several approaches to face this task have been experimented, and improvements in TE technologies have been shown in RTE evaluation campaigns, TE systems performances are still far from being optimal. Moreover, while systems developers create new modules, algorithms and resources to address specific inference types, it is difficult to measure a substantial impact when such modules are evaluated on RTE data sets because of *i)* the sparseness (i.e. low frequency) of the single phenomena, and *ii)* the impossibility to isolate each phenomenon, and to evaluate each module independently from the others.

A renewed interest is therefore rising in the TE community towards a deeper and better understanding of the core phenomena involved in textual inference, and a number of recently published works (Bentivogli *et al.* 2010 [11], Sammons *et al.* 2010 [83]) agree that incremental advances in local entailment phenomena are needed to increase the performances in the main task, which is perceived as omni comprehensive and not fully understood yet.

[1] `http://qallme.fbk.eu/`

1.3 The Solution

In line with the expectations of the TE community, we are convinced that crucial progress may derive from a focus on decomposing the complexity of the TE task into basic phenomena and on their combination. More specifically, basing on the original definition of TE, that allows to formulate textual inferences in an application independent way and to take advantage of available data sets for training provided in the RTE evaluation campaigns, the aim of our work is to analyse how the common intuition of decomposing TE would allow a better comprehension of the problem from both a linguistic and a computational viewpoint. We propose a framework for component-based TE, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. Five dimensions of the problem are investigated: *i)* the definition of a component-based TE architecture; *ii)* the implementation of system components able to address specific inference types; *iii)* the linguistic analysis of the phenomena relevant to component-based TE; *iv)* the automatic acquisition of knowledge to support component-based entailment judgements; *v)* the development of evaluation methodologies to assess component-based TE systems capabilities to address single phenomena in a pair.

1.4 Innovative Aspects/ Contributions

The first innovative aspect of this Thesis concerns the definition and implementation of a model to decompose the complexity of the Textual Entailment problem, assuming Fregean meaning compositionality principle. Starting with a study of the applied notion of Textual Entailment as outlined in the Computational Linguistics field, under the perspective of logical "argument", we compare TE pairs to certain categories of arguments,

and we evaluate them according to the criteria described in (Nolt *et al.* 1998 [75]). Taking advantage of those observations and definitions, we propose a model for TE pairs decomposition, to highlight the relations between the premise (i.e. T) and the conclusion (i.e. H). To benefit from this idea from a computational point of view, we have defined a framework for component-based TE, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. In this component-based architecture, a set of clearly identifiable TE modules can be singly used on specific entailment sub-problems and can be then combined to produce a global entailment judgement for a pair. Aspects related to meaning compositionality, which are absent in the original definition of TE, are introduced to bring new light into textual inference. To experiment the feasibility of the component-based TE framework described above, we implemented a set of TE-components basing on the architecture of the EDITS system (Kouylekov and Negri 2010 [48]). Even if such package was not developed whithin this Thesis work, we provided valuable contributions to its improvement, and we adapted its architecture to account for the properties of the TE-components we previously defined. We evaluated such components on RTE data sets, both *i)* independently - to test their precision to detect and solve the category of phenomena they are built to deal with - and *ii)* combining them using both linear and sequential composition models. Such architecture has been evaluated in our participations to RTE campaigns (in particular, RTE-4), on real RTE data sets provided by the organizers of the challenges, and in our participation EVALITA, where we carried out the TE task for Italian.

Important contributions of this Thesis are the pilot resources obtained as outcome of two different studies, the first one concerning the definition of a methodology for the creation of specialized data sets, made of atomic T-H pairs in which a certain phenomenon underlying the entail-

ment relation is highlighted and isolated, and the second one concerning the implementation of an algorithm for the acquisition of high precision entailment rules from Wikipedia revision history. More specifically, the first study resulted in the creation of two data sets,[2] made of *i)* 90 RTE-5 Test Set pairs (30 entailment, 30 contradiction and 30 unknown examples) annotated with linguistic phenomena relevant to inference (both with fine grained and macro categories), and *ii)* 203 atomic pairs created from the 90 annotated pairs (157 entailment, 33 contradiction, and 13 unknown examples). The second study, i.e. the implementation of an algorithm to automatically acquire knowledge in the form of entailment rules, was carried out on two experimental settings, to collect rules expressing causality and temporal expressions. The obtained resource[3] includes, respectively, 1249 and 665 rules, covering entailment and paraphrasing aspects not represented in other similar resources, and shows both high quality and coverage of the extracted rules. Since the methodology does not require human intervention, the resource can be easily extended and periodically updated, as Wikipedia revisions change continuously.

Finally, a further contribution of this research work is the development of a strategy to provide a more detailed evaluation of the capabilities of TE systems to address specific inference types. It takes advantage of the decomposition of T-H pairs into atomic pairs, and assumes that the more a system is able to correctly solve the linguistic phenomena underlying the entailment relation separately, the more the system should be able to correctly judge more complex pairs, in which different phenomena are present and interact in a complex way. As a pilot study, we have applied our evaluation methodology to the output of three systems that took part in RTE-5, i.e. EDITS, VENSES (Venice Semantic Evaluation System) and

[2]Available at `http://hlt.fbk.eu/en/Technology/TE_Specialized_Data`

[3]`http://hlt.fbk.eu/en/technology`

BLUE (Boeing Language Understanding Engine), and we discovered that, although the three systems have similar accuracy on RTE-5 data sets, they show significant differences in their respective abilities to manage different linguistic phenomena and to properly combine them. As an outcome, a more meaningful evaluation of RTE systems is provided, that highlights on which aspects a system needs to improve its performances, and the features it should focus on.

The list of the papers published in the course of the Doctoral School can be found in Appendix A.

1.5 Structure of the Thesis

The Thesis is structured as follows:

Chapter 2 analyses the applied notion of Textual Entailment as outlined in the Computational Linguistics field, under the perspective of logical "argument" as formulated in Philosophy of Language.

Chapter 3 presents the state of the art of the research in Textual Entailment, and the Recognizing Textual Entailment evaluation campaign. In particular, it focuses on the aspects of the works in the literature that are more relevant to the component-based framework for TE we propose in this Thesis.

Chapter 4 describes the framework for component-based Textual Entailment we propose, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. In this component-based architecture, a set of clearly identifiable TE modules can be singly used on specific entailment sub-problems and can be

then combined to produce a global entailment judgement for a pair. In particular, we propose a framework for the definition and combination of *transformation-based TE-components*, each of which able to deal with a certain aspect of language variability. We define them taking advantage of the conceptual and formal tools available from an extended model of Natural Logic (NL). Aspects related to meaning compositionality, which are absent in the original definition of TE, are introduced to bring new light into textual inference.

CHAPTER 5 describes the experimental work we carried out to prove the feasibility of the component-based TE framework. We take advantage of the modular architecture of the EDITS system (Edit Distance Textual Entailment Suite), an open-source software package for recognizing TE developed by the HLT group at FBK (Kouylekov and Negri 2010 [48]), and we used it as the basic architecture for the implementation of a set of TE-components. Strategies to combine the output of each single component in order to obtain a global entailment judgement for a pair are experimented.

CHAPTER 6 presents an analysis of the phenomena relevant to component-based TE. Moreover, it describes a methodology for the creation of specialized TE data sets made of *atomic T-H pairs*, i.e. pairs in which a certain phenomenon relevant to the entailment relation is highlighted and isolated, and describes the feasibility study we carried out applying the devised methodology to a sample of pairs extracted from the RTE-5 data set.

CHAPTER 7 presents an experimental strategy for the automatic acquisition of atomic T-H pairs and, in particular, of the entailment rules that allow to carry out the related inferential step. We take advantage of the syntactic structure of atomic pairs to define the more appropriate linguistic

constraints for the rule to be successfully applicable. We have carried out a large-scale application of our methodology on Wikipedia.

Chapter 8 introduces a new TE-systems evaluation, that takes advantage of the decomposition of Text-Hypothesis pairs into atomic pairs, and proposes to run systems over such data sets. As a result, a number of quantitative and qualitative indicators about strength and weaknesses of TE systems are highlighted.

Chapter 9 concludes the Thesis drawing final remarks and suggesting directions for future improvements.

Chapter 2

Semantic Inferences

Goal of this Chapter is to analyse the applied notion of Textual Entailment as outlined in the Computational Linguistics field, under the perspective of logical "argument" as formulated in Philosophy of Language.

2.1 Introduction

One of the essential activities carried out by humans in everyday linguistic interactions is the act of drawing a conclusion from given facts through some forms of reasoning. Given a sequence of statements (i.e. the premises), humans are (generally) able to *infer* or derive a conclusion that follows from the facts described in the premises. Since Aristotle, logicians and philosophers of language have developed theories to examine and formalize reasoning, that underlie the current attempts to emulate human inference developing automated systems aimed at natural language understanding.

Beside formal approaches to semantic inference that rely on logical representation of meaning, the notion of Textual Entailment (TE) has been proposed as an applied framework to capture major semantic inference needs across applications in the Computational Linguistics field.

Aim of this Chapter is to position and analyse the Textual Entailment

framework under the perspective of logical "argument", as formulated in Philosophy of Language. For this reason, we get back to the classical definition of *argument* (Section 2.2) and to the criteria outlined in logic to assess if an argument is a "good" argument, i.e. if it demonstrates the truth of its conclusion (Section 2.3). A classification of the types of semantic inference is provided in Section 2.4, to highlight the similarities of these forms of inductive reasoning with the kind of inferences addressed by TE. Then, we provide the classical definition of *entailment* in logic (Section 2.5), and a description of the traditional formal approaches to semantic inference (Section 2.6), discussing their limits in real world situations. Finally, in Section 2.8 we present the notion of *textual entailment*, and we analyse such applied framework adopting the definitions and the argument evaluation criteria formulated in logic. We point out discrepancies from a terminological viewpoint, since Textual Entailment seems to address both deductive and inductive arguments, the latter prevailing numerically on the first ones. Also issues related to the lack of a clear distinction between linguistic and world knowledge involved in the reasoning allowed by TE are discussed.

2.2 Logical argument

An *argument* is a sequence of statements of which one is intended as a *conclusion* and the others, the *premises*, are intended to prove or at least provide some evidence for the conclusion.[1] An example of a valid argument is given by the following well-known syllogism (2.1):

(2.1) *All men are mortal.*
Socrates is a man.
Therefore, Socrates is mortal.

[1]The definitions and the examples presented in this section and in Sections 2.3 and 2.4 are extracted from Nolt, Rohatyn and Varzi's manual of Logic [75].

In such argument, the first two statements are premises intended to prove the conclusion that Socrates is mortal. Premises and conclusion of an argument are always statements or propositions, i.e. assertions that can be either true or false, typically expressed by a declarative sentence (as opposed to questions, commands or exclamations). Though the premises must be intended to prove or provide evidence for the conclusion, it can be the case that some arguments are not too convincing, or are bad arguments. For instance, in Example 2.2 a child tries to persuade her mother to stay awake, but the fact that the movie is not over does not provide evidence to conclude that she cannot go to sleep. Logic aims therefore at developing methods and techniques to separate good arguments from the bad ones.

(2.2) *I can't go to bed, Mom. The movie's not over yet.*

Although the conclusion might occur either at the beginning or at the end in the argument, for purposes of analysis an argument is represented in its *standard form*, listing the premises in separate lines first, and then providing the conclusion (often marked with the symbol "$\therefore$", i.e. "therefore"). Example 2.3 shows the standard form of our previous example.

(2.3) *The movie's not over yet.*
$\therefore$ *I can't go to bed, Mom.*

Arguments occur only when someone intends a set of premises to support or prove a conclusion, and this intention is often expressed using peculiar words or phrases called *inference indicators*. They can be of two kinds:

- *conclusion indicators*, to highlight that the sentence is a conclusion from previously stated premises (e.g. *therefore*, *thus*, *as a result*);

- *premise indicators*, to signal that the sentence is a premise (e.g. *because*, *since*, *given that*).[2]

When placed between two propositions to form a compound sentence, such indicators are the main clues in identifying arguments and analysing their structure. For instance, given the following examples:

(2.4) *He is not at home, so he has gone to the movie.*

(2.5) *He is not at home, since he has gone to the movie.*

the inference indicators signal the reverse order premise-conclusion. In Example 2.4, "he has gone to the movie" is the conclusion, introduced by the indicator "so", while in Example 2.5 the same sentence is provided as the premise, because of the indicator "since". Some arguments do not have explicit indicators, and in order to differentiate premises from conclusions we must rely on the context or on our understanding of the author's intention.

In *complex arguments*, a conclusion is derived from a set of premises, and then that conclusion (also together with other statements) is used as a premise to draw a further conclusion, that may function as a premise for yet another conclusion, and so on. Those premises intended as conclusions from previous premises are called *nonbasic premises* or *intermediate conclusions*. For instance, given the following argument:[3]

(2.6) *All rational numbers are expressible as a ratio of integers. But pi is not expressible as a ratio of integers. Therefore pi is not a rational number. Yet clearly pi is a number. Thus there exists at least one nonrational number.*

All rational numbers are expressible as a ratio of integers.

[2]Some of these expressions have also other functions in different contexts, where no inference is assumed. For instance, "since" can indicate duration in *It has been six years since we went to France.*

[3]For a detailed analysis of the arguments reported in this sections, and for more examples, see Nolt *et al.* (1998) [75].

Pi is not expressible as a ratio of integers.
$\therefore$ *Pi is not a rational number.*

Pi is a number.
$\therefore$ *There exists at least one nonrational number.*

the first two premises support the intermediate conclusion that "Pi is not a rational number", that is in turn one of the premises to derive the final conclusion that "There exists at least one nonrational number". The complex argument described above is therefore made up of two steps of reasoning, that are arguments on their own right.

If an argument contains several steps of reasoning supporting all the same (final or intermediate) conclusion, the argument is said to be *convergent*, as in Example 2.7:

(2.7) *One should quit smoking. It is very unhealthy, and it is annoying to the bystanders.*

where the premises that smoking is unhealthy, and that it is an annoying action, are independent reasons to support the conclusion that one should quit smoking (i.e. each premise supports the conclusion separately). In other cases, a premise could instead require the support of other statements in order for the argument to make good sense, meaning that we need to assume the first premise to understand the step from the second premise to the conclusion. For instance, in Example 2.8 none of the premises (i.e. "Everyone at the party is a biochemist", "All biochemist are intelligent", and "Sally is at the party") taken separately would provide enough evidence to infer the conclusion that Sally is intelligent (the argument is not convergent).

(2.8) *Everyone at the party is a biochemist and all biochemists are intelligent. Therefore, since Sally is at the party, Sally is intelligent.*

Some arguments can be seen as incompletely expressed, and *implicit* premises (or conclusions) should be "read into" them, but only if they are required to complete the arguer's thought. For instance, Example 2.2 can be considered as incomplete, since the implicit premise "I can't go to bed until the movie is over" should be added to make it a good argument.[4] In some cases, the decision to regard the argument as having an implicit premise may depend on the degree of rigour which the context demands.

2.3 Argument evaluation

As introduced before, the main purpose of an argument is to demonstrate that a conclusion is true or at least likely to be true. It is therefore possible to judge an argument with respect to the fact that it accomplishes or fails to accomplish this purpose. In Nolt *et al.* (1998) [75], four criteria for making such judgements are examined: *i)* whether the premises are true; *ii)* whether the conclusion is at least probable, given the truth of the premises; *iii)* whether the premises are relevant to the conclusion; and *iv)* whether the conclusion is vulnerable to new evidence.[5]

2.3.1 Criterion 1: Truth of premises

The motivations for Criterion 1 are related to the fact that if any of the premises of an argument is false, it is not possible to establish the truth of its conclusion. Often the truth or falsity of one or more premises is unknown, so that the argument fails to establish its conclusion "so far as we know". In such cases, we may suspend the judgement until relevant

[4]To avoid misinterpretation, the argument should be made as strong as possible while remaining faithful to what one knows of the arguer's thought (*principle of charity*).

[5]Some of the proposed criteria are inapplicable to the arguments intended merely to show that a certain conclusion follows from a set of premises, whether or not the premises are true. However, in this chapter we are not concerned with these cases.

information that would allow us to correctly apply criterion 1 is acquired. Consider for instance Example 2.9, describing a situation where a window has been broken and a child tells us that she saw the person who broke it. In the standard format:

(2.9) *I saw Billy break the window*
$\therefore$ *Billy broke the window.*

Even if the child is telling the truth, her argument fails to establish its conclusion to us until we do not have evidence that the premise is true.

Criterion 1 is a necessary - but not sufficient - condition for establishing the conclusion, i.e. the truth of the premise does not guarantee that the conclusion is also true. In a good argument, the premises must adequately support the conclusion, and the criteria described in Sections 2.3.2 and 2.3.3 are thought to assess this aspect.

2.3.2 Criterion 2: Validity and inductive probability

The goal of criterion 2 is to evaluate the arguments with respect to the probability of the conclusion, given the truth of the premises. According to this parameter, arguments are classified into two categories:

- *deductive arguments*, whose conclusion follows *necessarily* from their basic premises (i.e. it is impossible for their conclusion to be false while the basic premises are true);
- *inductive arguments*, whose conclusion does not necessarily follow from their basic premises (i.e. there is a certain probability that the conclusion is true if the premises are, but there is also a probability that it is false).[6]

[6]In Nolt et al. (1998) [75], the authors highlight the fact that in the literature the distinction between inductive and deductive argument is not universal, and slightly different definitions can be found in some works.

Example 2.11 is a *valid* deductive argument[7] (as well as Example 2.1), while Example 2.12 has to be classified as an inductive argument.

(2.11) *No mortal can halt the passage of time.*
You are a mortal.
∴ *You cannot halt the passage of time.*

(2.12) *There are no reliably documented instances of human beings over 10 feet tall.*
∴ *There has never been a human being over 10 feet tall.*

Given a set of premises, the probability of a conclusion is called *inductive probability*, and it is measured on a scale from 0 to 1. The inductive probability of a deductive argument[8] is maximal, i.e. equal to 1, while the inductive probability of an inductive argument is (typically) less than 1.[9]

The fact that deductiveness and inductiveness are independent of the actual truth or falsity of the premises and conclusion (assessed by criterion 1) is clearly evident in Example 2.13, where all the statements are false.

(2.13) *Some pigs have wings.*
All winged things sing.
∴ *Some pigs sing.*

In an inductive or a deductive argument, any combination of truth or falsity is possible, except that no deductive (valid) argument ever has true

[7] *Invalid* deductive arguments are arguments which claim to be deductive, but in fact are not, as:

(2.10) *Some Greeks are logicians.*
Some logicians are tiresome.
∴ *Some Greeks are tiresome.*

Example 2.10 is an invalid argument, because, e.g. the tiresome logicians might all be Romans. Arguments can be invalid for a variety of reasons, due to misunderstanding or misinterpretation during the reasoning process on the premises (see Chapter 8 of Nolt *et al.* 1998 [75] for a more exhaustive classification of fallacies).

[8] From here on, with the term *deductive argument* we refer to valid deductive arguments only.

[9] In this Chapter, we will not discuss some controversial theories of inductive logic on the value of the inductive probability of an inductive argument. For further details, see Carnap (1962) [19].

premises and a false conclusion (by definition it is impossible[10] that in a deductive argument a false conclusion follows from true premises). A deductive argument is said to be *sound* if all its premises are true (as, for instance, Example 2.11).

Although deductive arguments provide the greatest certainty (inductive probability = 1), in practice we must often settle for inductive reasoning, that allows for a range of inductive probabilities and varies widely in reliability. When the inductive probability of an inductive argument is high, the reasoning of the argument is said to be *strong* or *strongly inductive*. On the contrary, it is said to be *weak* or *weakly inductive* when the inductive probability is low. There is no clear distinction line between strong and weak inductive reasoning, since these definitions can be context-dependent (in general an argument is weak if its inductive probability < 0.5). Furthermore, since only in a few cases the information contained in the statements of inductive arguments is numerically quantifiable, often it is not possible to provide a precise number as the inductive probability.

In complex arguments (introduced in Section 2.2), the deductive validity and inductive probability are relations between the basic premises and the conclusion. Each of the steps that make up a complex argument is in itself an argument, and has its own inductive probability. Assessing how such inductive probabilities correlate with the inductive probability of the complex argument to which they belong is not an easy task. In Nolt *et al.* (1998) [75] the authors suggest some rules of thumb: *i)* in complex nonconvergent arguments, if one or more of the steps are weak, then usually the inductive probability of the argument as a whole is low; *ii)* if all the steps of a complex nonconvergent argument are strongly inductive or deductive, then (if there are not too many of them) the inductive probability of the whole is fairly high; *iii)* the inductive probability of a

[10]It means *logically impossible*, i.e. impossible in its very conception.

convergent argument is usually at least as high as the inductive probability of its strongest branch. Since all these rules allow for some exceptions, the only way to obtain an accurate judgement of the inductive probability of the arguments described in the rules is to examine directly the probability of the conclusion given the basic premises, and ignoring the intermediate steps.

2.3.3 Criterion 3: Relevance

Criterion 3 claims that any argument which lacks relevance (regardless of its inductive probability) is useless for demonstrating the truth of its conclusion (it is said to commit a *fallacy of relevance*). For instance, the premises of the arguments shown in Examples 2.9 and 2.11 are highly relevant to derive their conclusion.

Relevance and inductive probability do not always vary together, i.e. some arguments can be strongly inductive with low relevance, or weakly inductive with high relevance. The first type is represented by arguments whose conclusions are logically necessary (e.g. tautologies, as *No smoker is a nonsmoker*), and therefore true under any conditions (such arguments are deductive by definition). Another case of deductive arguments with low relevance occurs when the premises are inconsistent, i.e. they cannot all be true simultaneously, as in Example 2.14. By definition, any argument with inconsistent premises is deductive regardless of its conclusion.[11]

(2.14) *All butterflies are insects.*
Some butterflies are not insects.

[11] For a more detailed explanation, see Nolt *et al.* (1998) [75].

2.3.4 Criterion 4: The requirement of total evidence

One of the most important differences between inductive and deductive arguments concerns their vulnerability to new evidence, meaning that while adding new premises to deductive arguments make them remain deductive, the inductive probability of inductive arguments can be strengthened or weakened by the introduction of new information. For instance, the argument showed in Example 2.15 is strongly inductive:

(2.15) *Very few Russians speak English well.*

Sergei is Russian.

$\therefore$ *Sergei does not speak English well.*

but if the following premises are added:

(2.16) *Sergei is an exchange student at an American university.*

Exchange students at American universities almost always speak English well.

the inductive probability is reduced (the new premises provide evidence against the conclusion supported by the first two statements). The choice of the premises in an inductive reasoning is therefore crucial, since a conclusion may appear as more or less probable according to the evidences selected to support it.

For this reason, the criterion of *total evidence condition* stipulates that if an argument is inductive its premises must contain all known evidence that is relevant to the conclusion. Inductive arguments which fail to meet this requirement are said to commit the *fallacy of suppressed evidence*, that can be committed either intentionally or unintentionally.

2.4 Inductive reasoning

The inductive probability of an inductive argument depends on the relative strengths of its premises and conclusion. Nolt *et al.* (1998) [75] claim that the strength of a statement is determined by what the statement says, i.e. the more it says, the stronger it is (regardless of the truth of its content). The truth of a strong statement is proved only under specific circumstances, while since the content of a weak statement is less specific, its truth can be verified under a wider variety of possible circumstances. For these reasons, the strength of a statement is approximately inversely related to its *a priori* probability, i.e. the probability prior or in the absence of evidence: the stronger the statement is, the less inherently likely it is to be true, while the weaker it is, the more probable it is. Let's consider the following examples:

(2.17) *Some people are sort of weird.*

(2.18) *Every vertebrate has a heart.*

While Example 2.17 is a weak statement because it says nothing very specific (i.e. some people are weird but it can be the case that some other are not), Example 2.18 is a strong statement because it asserts that all the vertebrates have a certain characteristic (i.e. there cannot exist a vertebrate without a heart).

It is not always straightforward to compare the strengths of the statements, and in order to rank some sets of statements with respect to their relative strength some rules must be followed: *i)* if statement A deductively implies statement B, but B does not deductively imply A, then A is stronger than B (i.e. the circumstances in which A is true are a subset of the possible circumstances in which B is true); *ii)* if statement A is logically equivalent to statement B (i.e. if A and B deductively imply one

another), than A and B are equal in strength. However, such rules are not always applicable, and sometimes the differences in strength among a set of statements are too small to be intuitively apparent.

The concept of strength of a statement has been introduced here because of its relation to inductive probability, since the latter tends to vary directly with the strength of the premises, and inversely with the strength of the conclusion. For instance, in Example 2.19 the premise gets stronger as the number n gets larger, and the argument's inductive probability increases as well.

(2.19) *We have observed at least n daisies, and they have all had yellow centers.*
$\therefore$ *If we observe another daisy, it will have a yellow center.*

Inductive arguments can be divided into two types: *i)* the *Humeian* arguments (after the philosopher David Hume who was the first to study them) require the presupposition that the universe or some aspect of it is or is likely to be uniform or lawlike (we will discuss them in Sections 2.4.3, 2.4.4 and 2.4.5); and *ii)* the *statistical* arguments, which do not require this presupposition, and the conclusions are supported by the premises for statistical or mathematical reasons (Sections 2.4.1 and 2.4.2).

2.4.1 Statistical syllogism

Statistical syllogism is an inference from statistics concerning a set of individuals, to a (probable) conclusion about some members of that set. According to the *logical* interpretation of the inductive probability, its value in a statistical argument is the percentage figure divided by 100.[12] For instance, in Example 2.20 the inductive probability is 0.98.

(2.20) *98% of college freshmen can read beyond the 6th-grade level.*

[12]According to the *subjective* interpretation, the inductive probability is a measure of a particular rational person's degree of belief in the conclusion, given the premises.

Dave is a college freshman.
$\therefore$ *Dave can read beyond the 6th-grade level.*

The argument in Example 2.20 is called *statistical syllogism*, and can be formalized as:

n% of F are G.
x is F.
$\therefore$ *x is G.*

where F and G should be replaced by predicates, x by a name and n by a number from 0 to 100. As introduced before, the inductive probability of a statistical syllogism is $n/100$; if $n = 100$ the argument is deductive, while if $n < 50$ the form of the argument becomes the following:

n% of F are G.
x is F.
$\therefore$ *x is not G.*

and its inductive probability is $1 - n/100$ (in this case, if $n = 0$, the argument is deductive).

A precise inductive probability cannot be assigned to arguments whose premises do not provide numerical values as statistics, as Example 2.21.

(2.21) *Madame Plodsky's diagnoses are almost always right.*
Madame Plodsky says that Susan is suffering from a kidney stone.
$\therefore$ *Susan is suffering from a kidney stone.*

Anyway, as explained in Section 2.3, other criteria should be considered in argument evaluation. For instance, Example 2.21 is an argument from authority, whose strengths depend on Mme Plodsky's reliability (if Mme Plodsky is a fortune teller, the first premise is maybe false). If the first premise "Madame Plodsky's diagnoses are almost always right" is omitted,

the argument is no longer a statistical syllogism: the remaining premise lacks relevance to the conclusion, since the evidence that the authority is reliable is missing (its inductive probability drops significantly).[13]

Let's replace the first premise of Example 2.21 with the statement "Most of Mme Plodsky's diagnosis are false". This new argument is a form of *ad hominem* argument, that reasons from the unreliability of a person's pronouncement.

2.4.2 Statistical generalization

Statistical generalization is an inference from statistics related to a randomly selected subset of a set of individuals, to a (probable) conclusion about the composition of the set as a whole, as shown in Example 2.22.

(2.22) *50% of 1000 randomly selected Americans said that they support Obama.*
∴ About 50% of all Americans would say (if asked under the survey conditions) that they support Obama.

The general form of such kind of inductive reasoning is the following:

n% of s randomly selected F are G.
∴ About n% of all F are G.

where s is the size of the sample, F is a property defining the population about which we are generalizing (the Americans, in Example 2.22), and G is the property studied by the survey (in this case, the property of supporting U.S. President Obama). The sample from which we are generalizing should be *i)* randomly selected, so that each of the F's had the same chance of being sampled; *ii)* fairly large. If the sample is not randomly chosen, it is said to be *biased*, and attempts to apply statistical generalization on a

[13] A *fallacy of appeal to authority* is committed (more details can be found in Nolt *et al.* 1998 [75]).

biased sample commit the *fallacy of biased sample* (a form of the *fallacy of hasty generalization*).

The inductive probability of a statistical generalization is calculated basing on mathematical principles, and is a function of the sample size (the bigger the size, the stronger the premises) and the strength of the conclusion (we must allow it a certain margin of error, so terms like *about* provide more reliability).[14] If the conclusion is too strong to be supported with reasonable inductive probability by the premises, the argument is said to commit the *fallacy of small sample* (another form of the *fallacy of hasty generalization*).

2.4.3 Inductive generalization and simple induction

Often it is not possible to obtain a random sample of the population on which we want to focus our study, e.g. if it concerns future objects or events. For instance, the conclusion of the argument in Example 2.23 considers all the games played by the Bat this season, which include future games:

(2.23) *The Bats won 10 out of 20 games they have played so far this season.*
∴ The Bat will finish the season having won about half of their games.

This kind of inductive reasoning is called *inductive generalization*, and its general form can be represented as follows:

n% of s thus-far-observed F are G.
∴ About n% of all F are G.

[14]Mathematical methods can be used to calculate the argument's inductive probability numerically, if this margin of error is delineated precisely. As a result we could, for instance, replace the conclusion of Example 2.22 with the following statement: "50% ± 10% of all Americans would say (if asked under the survey conditions) that they support Obama".

Differently from statistical generalization, the premise in the inductive generalization does not claim that the sample is random, so mathematical principles cannot justify the reasoning (no mathematical principle can guarantee the results of the next games played by the Bats). Inductive generalizations are Humean inferences, since they presuppose that the course of the events exhibits or is likely to exhibit a certain uniformity over time (as in Example 2.23). Inductive generalizations are weaker arguments than statistical generalization, but their evaluations are based on the same principles (in both cases, inductive probability increases as s does).
When $n = 100$, the general form of the inductive generalization becomes:

All the s thus-far-observed F are G.
$\therefore$ *All F are G.*

and represents the form by which scientific laws are justified.

Reducing the population on which the argument focuses to one individual is the most extreme way to weaken the conclusion. This is represented by the following form, called *simple induction*, *induction by enumeration*, or *the simple predictive inference*:

n% of the s thus-far-observed F are G.
$\therefore$ *If one more F is observed, it will be G.*

Simple inductions are generally stronger than inductive generalization from the same premises.

2.4.4 Induction by analogy

Arguments by analogy are another kind of Humean arguments. In these arguments we observe that an object x has many properties, $F_1, F_2, ..., F_n$ in common with some other object y. We observe also that y has some

further property G. Hence, we consider it likely (since x and y are analogous in so many other respects) that x has G as well, as shown in Example 2.24.

(2.24) *Specimen x is a single-stemmed plant with lanceolate leaves and five-petals blue flowers, about 0.4 meter tall, found growing on a sunny roadside.*
Specimen y is a single-stemmed plant with lanceolate leaves and five-petals blue flowers, about 0.4 meter tall, found growing on a sunny roadside.
Specimen y is a member of the gentian family.
$\therefore$ *Specimen x is a member of the gentian family.*

The general form of an argument by analogy can be represented as follows:

F_1x & F_2x & ... & F_nx
F_1y & F_2y & ... & F_ny
Gy
$\therefore Gx$

Like other kinds of inductive arguments, analogical arguments can be strengthened by strengthening their premises (i.e. adding more properties that x and y have in common) or by weakening their conclusions (e.g. replacing the conclusion of Example 2.24 with the statement "Specimen x is a member of the gentian family or some closely related family"). It must be noted, however, that the strength of the premises does not depend only on the number of the properties that x and y have in common, but also on the specificity of these properties, and the relevance of the properties to G. As all inductive arguments, analogical arguments are vulnerable to contrary evidence, that often takes the form of a *relevant disanalogy*.

2.4.5 Causality

To determine the cause of an observed effect, usually humans carry out a two-step procedure: first, they formulate a list of the suspected causes, and

then by observation they rule out the highest number of these suspected causes to conclude that the item left is the likely cause of the effect. Since the first step is generally inductive (frequently it is an analogical reasoning), while the eliminative reasoning of the second step is deductive, the reasoning as a whole is considered to be inductive.

In Nolt *et al.* (1998) [75], four different kinds of causes are listed:

- *necessary cause* or *causally necessary condition*: a necessary cause for an effect E is a condition which is needed to produce E. If C is a necessary cause for E, then E will never occur without C, though perhaps C can occur without E. A given effect can have several necessary causes, e.g. to produce fire, three causally necessary conditions are needed: fuel, oxygen and heat;

- *sufficient cause* or *causally sufficient condition*: a condition C is a sufficient cause for an effect E, if the presence of C invariably produces E. If C is a sufficient cause for E, then C will never occur without E, though there may be cases in which E occurs without C (e.g. decapitation is a sufficient cause for death, but the converse does not hold). A given effect can have several necessary causes;

- *necessary and sufficient causes*: the effect E never occurs without the cause C, nor the cause C without the effect E (e.g. the presence of a mass is causally necessary and sufficient for the presence of a gravitational field: no mass, no gravitational field);

- *causal dependence of one variable quantity on another*: a variable quantity B is causally dependent on a second variable quantity A, if a change in A always produces a corresponding change in B (e.g. raising the temperature of a gas will cause an increase in its volume).

Correspondingly to each cause, a different method of elimination has been investigated by the philosopher John Stuard Mill:[15]

- *method of agreement* is a deductive procedure for ruling out suspected causally necessary conditions. As introduced before, if a circumstance C is a causally necessary condition of an effect E, then E cannot occur without C. So to determine which of a list of suspected causally necessary conditions really is causally necessary for E, a number of different cases of E should be examined. If any of the conditions fails to occur in any of these cases, then it can certainly be ruled out as not necessary for E.

- *method of difference* is a method for ruling out suspected causally sufficient conditions. As introduced before, a sufficient cause for an effect E is an event that always produces E. If cause C ever occurs without E, then C is not sufficient for E: any item of the list which occurs without E should be rejected. Claims of causal sufficiency are often implicitly to be understood as relative to a particular class of individuals or events.

- *method of agreement and difference* is a procedure for ruling out suspected necessary and sufficient conditions. It involves the simultaneous application of the methods of agreement and difference. If C is a necessary and sufficient cause of E, then C never occurs without E and E never occurs without C. Hence, in any case in which C occurs but E does not, or E occurs but C does not, C can be ruled out as a necessary and sufficient cause of E.

- *method of concomitant variation* does not concern the mere presence or absence of cause and effect, but their relative magnitude. Its goal

[15] A more detailed explanation of the methods and examples can be found in Nolt *et al.* (1998) [75].

is to narrow down a list of variable magnitudes suspected of being responsible for a specific change in the magnitude of an effect E. If that variable remains constant throughout the change, it is rejected as not responsible for that specific change. If all but one of a list of variables remain constant while the magnitude of an effect changes, and presuming that the variable responsible for the change appears on the list, it must the one which has not remained constant.

2.5 The notion of Entailment

As highlighted in the previous sections, if an argument is deductively valid, one should be able to infer or derive the conclusion from the premises, i.e. to show how the conclusion actually follows from the premises (Nolt *et al.* 1998 [75]). More specifically, a set of premises is said to *entail* a conclusion if the premises deductively imply the conclusion and in addition are relevant to it.

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 shown in Figure 2.1.

Ultimately, we want to regard entailment as a relation between utterances (that is, sentences in context), where the context is relevant to understand the meaning. In (Chierchia and McConnell-Ginet 2000 [21]), entailment is defined as a relation between sentences (S and S'), and the previous definition is simplified as: S entails S' iff whenever S is true, also S' is.

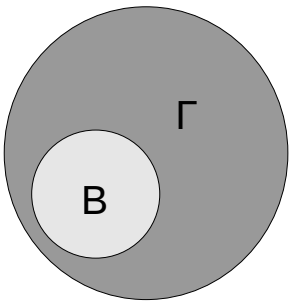


Figure 2.1: Venn diagram of the entailment relation

2.6 Computational approaches to semantic inference

Classical approaches to semantic inference rely on logical representations of meaning that are external to the language itself, and are typically independent of the structure of any particular natural language. Texts are first translated, or interpreted, into some logical form and then new propositions are inferred from interpreted texts by a logical theorem prover.

While propositional logic deals with simple declarative propositions, first-order logic additionally covers predicates and quantification. For instance, given the axiom "All greedy kings are evil" (Russel and Norvig 2002 [82]), formalized as:

(2.25) $\forall x \; King(x) \wedge Greedy(x) \Rightarrow Evil(x)$

it seems quite permissible to infer any of the following sentences:

$$King(John) \wedge Greedy(John) \Rightarrow Evil(John)$$
$$King(Richard) \wedge Greedy(Richard) \Rightarrow Evil(Richard)$$
$$King(Father(John)) \wedge Greedy(Father(John)) \Rightarrow Evil(Father(John))$$

A step-by-step deduction reasoning is performed applying a set of *rules of inference*, that allow to reach the conclusion through a finite number of successive steps of reasoning, each of which is fully explicit and indisputable

(Russel and Norvig 2002 [82], Nolt *et al.* 1998 [75]). Many deductive systems for first-order logic have been developed, showing both *soundness* (i.e. only correct results are derived) and *completeness* (i.e. any logically valid implication is derived).

But, especially after the development of the web, we have witnessed a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. Addressing the inference task by means of logical theorem prover in automated applications aimed at natural language understanding has shown several intrinsic limitations (Blackburn *et al.* 2001 [15]). As highlighted by Monz and de Rijke (2001) [70], in formal approaches semanticists generally opt for rich (i.e. including at least first order logic) representation formalisms to capture as many relevant aspects of the meaning as possible, but practicable methods for generating such representations are very rare. The translation of real-world sentences into logic is difficult because of issues such as ambiguity or vagueness (Pinkal 1995 [79]). Furthermore, the computational costs of deploying first-order logic theorem prover tools in real world situations may be prohibitive, and huge amounts of additional knowledge are required. The type of additional knowledge that can be needed ranges from linguistic knowledge, e.g. about word meaning, to non-linguistic background knowledge.

Formal approaches address forms of deductive reasoning, and therefore often exhibit a too high level of precision and strictness as compared to human judgements, that allow for uncertainties typical of inductive reasoning (Bos and Markert 2006 [16]). While it is possible to model elementary inferences on the precise level allowed by deductive systems, many pragmatic aspects that play a role in everyday inference cannot be accounted for. Inferences that are plausible but not logically stringent cannot be modelled in a straightforward way, but in NLP applications approximate reasoning should be preferred in some cases to having no answers at all. Espe-

cially in data-driven approaches, where patterns are learnt from large-scale naturally-occurring data, we can settle for approximate answers provided by efficient and robust systems, even at the price of logic unsoundness or incompleteness. Starting from these considerations, Monz and de Rijke (2001) [70] propose to address the inference task directly at the textual level instead, exploiting currently available NLP techniques. In [70], they experiment a method for entailment checking based on a similarity measure from information retrieval, sketching the framework that will be on the grounds of the operational definition of entailment described in Section 2.8.

2.7 Semantic inferences and language variability

While methods for automated deduction assume that the arguments in input are already expressed in some formal meaning representation (e.g. first order logic), addressing the inference task at a textual level opens different and new challenges from those encountered in formal deduction. Indeed, more emphasis should be put on informal reasoning, lexical semantic knowledge, and variability of linguistic expressions. To some extent, the problem of natural language inference moves away from earlier studies on logical inference, and becomes a central topic in Natural Language Processing. Issues such as syntactic parsing, morphological analysis, word sense disambiguation and lexical semantic relatedness, which were absent in the previous scenario, become essential elements of this new framework. To identify implications in natural language sentences, automatic systems are therefore asked to deal with different linguistic phenomena and with a broad variety of semantic expressions. Indeed, language variability manifests itself at different levels of complexity, and involves almost all linguistic phenomena of natural languages, including lexical, syntactic and semantic

variations. As an example, let's consider the following textual snippets:

(2.26) a. *Opposition supporters threw rocks during rioting with pro-Mubarak supporters near Tahrir Square in Cairo.*

b. *Opponents of Egypt's President Mubarak go on the offensive, pushing counter-demonstrators out of side streets around Cairo's Tahrir square.*

c. *Egyptian anti-government protesters have fought back against supporters of President Hosni Mubarak, pushing them out of some streets near Cairo's Tahrir Square.*

The three textual fragments of Example 2.26 are extracted from today's newspapers, and all describe the riotous event that took place recently in Egypt against the current government. Here, lexical variability is expressed by the use of the synonyms *opponent* - *protesters*, while syntactic variability comes out from the use of different syntactic constructions as in the genitive/prepositional alternation *Cairo's* - *in Cairo*. Variability concerns also discourse phenomena, meaning that in order to recognize that the event described involves the same entities, *them* must be referred to the correct entity *supporters*, and *Mubarak* must be recognized as *President Hosni Mubarak*. The most frequent type of language variability is the semantic one, that requires to perform some reasoning about the meaning of words and world knowledge in order to derive certain information from the text. For instance, the same relation between two entities can be expressed with event or relation in a cause/effect alternation, as in a) *threw rocks during rioting* - b) *go on the offensive* - c) *have fought back against*. Furthermore, when humans read a text, they derive meanings exploiting their knowledge about the world. Semantic variation in the text snippets of Examples 2.26 is recognized basing on our common knowledge that being part of the *opposition* movement implies supporting the *anti-government* party and so on. Other types of semantic variability are connected with

temporal and numerical expressions, requiring for instance the ability to reason about time and space, as in Example 2.27.

(2.27) a. *Apollo 14 landed on the Moon 40 years ago this week.*

b. *Apollo 14 landed on the Moon in 1971.*

Natural language inference systems should therefore exploit the achievements reached in NLP tasks such as syntactic parsing, computational lexical semantics and coreference resolution, in order to tackle the more challenging problems of sentence-level semantics.

2.8 Textual Entailment

As a generic framework for modelling language variability and capturing major semantic inference needs across applications in NLP, Dagan and Glickman (2004) [28] propose the notion of Textual Entailment. It is defined 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, since the meaning of one expression may usually entail the other, while entailment in the other direction is much 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 and Malakasiotis 2010 [3], Dagan *et al.* 2009 [27]). For instance, a question answering (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 information retrieval (IR) relevant retrieved documents should entail the combination of semantic concepts and relations denoted by the query. In information extraction (IE), entailment holds between different text variants expressing the same target relation (Romano *et al.* 2006 [81]). In text summarization (SUM), 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 and McKeown 2005 [9]), 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 Machine Translation (MT), an entailment relation should hold *i)* among machine-generated translations and human-authored ones that may use different phrasings in the evaluation phase (Pado *et al.* 2009 [76]), or *ii)* in the translation phase, between source language words and longer phrases that have not been encountered in training corpora (Mirkin *et al.* 2009 [68]). Other applications that could benefit from such inference model are reading comprehension systems (Nielsen *et al.* 2009 [74]).

While entailment in its logical definition pertains to the meaning of language expressions, in this applied model inferences are performed directly over lexical-syntactic representations, as typically obtained from syntactic parsing. Differently from the classical semantic definition of entailment provided in (Chierchia and McConnell-Ginet 2000 [21]) and discussed in Section 2.5, the notion of textual entailment accounts for some degree of uncertainty allowed in applications, as shown in Examples 2.28 and 2.29:

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

H: *Coffee drinking has health benefits.*

(2.29) T: *The technological triumph known as GPS was incubated in the mind of Ivan Getting.*

H: *Ivan Getting invented the GPS.*

In these cases, the truth of the hypothesis is highly plausible, rather than certain, but we would expect them to be considered as good examples of inferences in text-based applications.

2.8.1 Probabilistic Textual Entailment

Glickman *et al.* (2006) [42] 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. Going back to the discussions on argument evaluation criteria presented in Section 2.3, this applied definition of entailment seems to be closer to the notion of inductive argument than to the definition of deductive argument, (almost) equivalent to the classical definition of entailment. For instance, according to the criterion of validity described in Section 2.3, Example 2.30 (argument standard format of Example 2.28, where T is decomposed in set of premises and H is the conclusion) would be evaluated as an inductive argument with a high inductive probability.

(2.30) *Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake.*

Consuming coffee apparently also can help reduce the risk of diseases.

$\therefore$ *Coffee drinking has health benefits.*

Also in (Zaenen *et al.* 2005 [98]) closely related issues are discussed (i.e. the relation between TE and classical notions such as presuppositions and implicature), and for these reasons they propose to refer to such a relation as *textual inference*, rather than textual entailment (see also Manning 2006 [58]).

2.8.2 TE and background knowledge

As introduced before, TE definition is based on (and assumes) common human understanding of language, as well as common background knowledge. However, the entailment relation is said to hold only if the statement in the text licenses the statement in the hypothesis, meaning that the content of T and common knowledge together should entail H, and not background knowledge alone. For this reason, in Example 2.31 T does not entail H.

(2.31) T: *Excessive amounts of pesticides and chemical fertilizers may be poisoning huge tracts in India.*

H: *Pesticides ruin fruits.*

With this respect, instead of viewing a T-H pair as true of false entailment, we agree with Manning (2006) [58] that it would be more appropriate to say if the hypothesis "follows" or "does not follow" from the text, somehow referring to the criterion of relevance discussed in Section 2.3. At the same time, what we assume as background knowledge to be introduced in the inference process in not completely clear. In their discussion, Dagan *et al.* (2006) [30] say that the criteria defining what constitutes acceptable background knowledge may be hypothesis dependent, and referring to Example 2.32 they claim that it is inappropriate to assume as background knowledge that the national language of Yemen is Arabic, since this is exactly the hypothesis in question. On the other hand, they claim that such

background knowledge might be assumed when examining the entailment "Grew up in Yemen" "Speaks Arabic".

(2.32) T: *The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.*

H: *The national language of Yemen is Arabic.*

Still, such clarification seems to us quite vague, and we agree with Manning (2006) [58] on the consideration that the amount of common-sense and general world knowledge is vast and not easily delineated, and it is much easy to stick with saying that world knowledge is "things that most people know". Also Zaenen *et al.* (2005) [98] discuss on the role of world knowledge in the inference task, and even if initially they say that they do not accept any, they gradually admit that it is impossible to filter it out. Furthermore, as Manning (2006) [58] claims, since TE aims at capturing the inference needs of applications aimed at natural language understanding, it would be wrong to exclude common sense and basic world knowledge.

2.8.3 Applying argument evaluation criteria to TE pairs

Bearing in mind the critical issues related to the notion of TE discussed in the previous section, let's try to judge some T-H pairs[16] with respect to the argument evaluation criteria described in Section 2.3. In general, in TE we assume the fact that if T and H refer to an entity x, the entity meaning is the same. First of all, we represent the entailment pair in the standard format of a logic argument, where T is a (set of) premise(s), and H the conclusion that must be inferred from the premises.

(2.33) a. T: *In 1541 the Turks took the Buda and held it until 1686; the city changed very little during this time.*

[16]Extracted from the data sets provided by the organizers of the Recognizing Textual Entailment (RTE) challenge, described in Chapter 3.

H: *The Turks held the Buda between 1451 and 1686.*

b. *In 1541 the Turks took the Buda.*
The Turks held the Buda until 1686.
The city changed very little during this time.
$\therefore$ *The Turks held the Buda between 1541 and 1686.*

c. **Criterion 1 - truth of premises:** at first, we suspend the judgement due to a lack of knowledge about the event described in the premises. After collecting new evidence (i.e. checking the information about the siege of Buda in the Internet, or in an encyclopaedia), the truth of the premises is verified.
Criterion 2 - validity and inductive probability: valid deductive argument.
Criterion 3 - relevance: the first two premises are relevant, while the third one is not relevant to infer the conclusion.
Criterion 4 - total evidence condition: deductive arguments are not vulnerable to new evidences.

Example 2.33 is a valid deductive convergent argument (i.e. it is necessary to assume the evidence provided by both premises to infer the conclusion), while the argument shown in Example 2.34 is based on a strong inductive reasoning.

(2.34) a. T: *The Crathes castle served as the ancestral seat of the Burnetts of Leys until gifted to the National Trust for Scotland by the 13th Baronet of Leys, Sir James Burnett in 1951.*
H: *Sir James Burnett was the owner of the Crathes castle.*

b. *The Crathes castle served as the ancestral seat of the Burnetts of Leys.*
The 13th Baronet of Leys, Sir James Burnett, gifted the Crathes castle

to the National Trust for Scotland in 1951.
∴ Sir James Burnett was the owner of the Crathes castle.

c. **Criterion 1 - truth of premises:** at first, we suspend the judgement due to a lack of knowledge about the event described in the premises. After collecting new evidence (i.e. checking the information about the owner of the Crathes castle in the Internet, or in an encyclopaedia), the truth of the premises is verified.
Criterion 2 - validity and inductive probability: inductive argument, high inductive probability.
Criterion 3 - relevance: satisfied.
Criterion 4 - total evidence condition: satisfied, as far as we know.[17]

On the contrary, Example 2.35 satisfies the first two criteria, but the premises do not provide any evidence to infer the hypothesis's truth. We can say that this argument commits the fallacy of suppressed evidence, since it does not provide any information concerning the place where the meeting took place.

(2.35) a. T: *Mr. Guido di Tella, Argentine foreign minister, met representatives of British companies and financial institutions.*
H: *Foreign Minister Guido De Tella went to the UK.*

b. *Mr. Guido di Tella is Argentine foreign minister.*
Mr. Guido di Tella met representatives of British companies and financial institutions.
∴ Foreign Minister Guido De Tella went to the UK.

c. **Criterion 1 - truth of premises:** at first, we suspend the judgement due to a lack of knowledge about the event described in the premises. After collecting new evidence (i.e. checking the information about Mr. di Tella in the Internet, or in an encyclopaedia), the truth of the premises is verified.

[17] Actually, premises claiming that Sir James Burnett was disinherited due to some reasons could bring new evidence that would contradict the conclusion, but we consider it to be not very likely.

Criterion 2 - validity and inductive probability: inductive argument, quite high inductive probability.

Criterion 3 - relevance: not satisfied, the text does not contain enough information to infer the hypothesis truth.

Criterion 4 - total evidence condition: not satisfied, vulnerable to the addition of new evidence (e.g. "The meeting took place in London/ Buenos Aires").

Example 2.36 is an invalid argument (the conclusion contradicts the premises).

(2.36) a. T: *The Communist Party USA was a small Maoist political party which was founded in 1965 by members of the Communist Party around Michael Laski who took the side of China in the Sino-Soviet split.*
H: *Michael Laski was an opponent of China.*

b. *The Communist Party USA was a small Maoist political party.*
The Communist Party USA was founded in 1965 by members of the Communist Party around Michael Laski.
Michael Laski took the side of China in the Sino-Soviet split.
$\therefore$ *Michael Laski was an opponent of China.*

c. **Criterion 1 - truth of premises:** at first, we suspend the judgement due to a lack of knowledge about the event described in the premises. After collecting new evidence (i.e. checking the information about the Communist Party USA and Michael Laski in the Internet, or in an encyclopaedia), the truth of the premises is verified.
Criterion 2 - validity and inductive probability: invalid argument (contradiction).
Criterion 3 - relevance: satisfied.
Criterion 4 - total evidence condition: not relevant.

As can be seen from these examples, in most of the cases entailment pairs have more in common with the inductive arguments described in Section

2.4, than with deductive arguments. However, applying the criteria of argument evaluation to T-H pairs is not always an easy task, and the issue related to the amount of common background knowledge we allow in our inference process can strongly affect our judgement on a certain argument. Consider Example 2.37:

(2.37) a. T: *Regan attended a ceremony in Washington to commemorate the landings in Normandy.*

H: *Washington is located in Normandy.*

b. *Regan attended a ceremony in Washington to commemorate the landings in Normandy.*

$\therefore$ *Washington is located in Normandy.*

c. **Criterion 1 - truth of premises:** at first, we suspend the judgement due to a lack of knowledge about the event described in the premises. After collecting new evidence (i.e. checking the information about the travels of President Regan in Internet, or in an encyclopaedia), the truth of the premises is verified.

Criterion 2 - validity and inductive probability: without considering background knowledge, it is a quite strong inductive reasoning. Somehow, we would infer that if a ceremony is held in a certain town x to commemorate something happened in a certain region y, x is located in y (e.g. if Normandy was replaced by USA, that inference would have been plausible).

Criterion 3 - relevance: premises are relevant to infer the conclusion.

Criterion 4 - total evidence condition: it commits the fallacy of suppressed evidence.

This argument commits (unintentionally?) the fallacy of suppressed evidence, i.e. some information is omitted in the premises due to lack of knowledge. To correctly evaluate Example 2.37 as invalid argument, the implicit premise "Washington is located in the U.S." should be added.

Since NLP applications are expected to correctly perform this kind of reasoning, "static" background knowledge should be extracted from external resources or knowledge-bases, and used in the inference process to convey new evidence to strengthen or invalid the reasoning.

Since argument evaluation criteria are applied by humans, reasoning is somehow performed at a high level, meaning that the problem of language variability discussed in Section 2.7 is not taken into consideration: it is part of the linguistic knowledge of a language owned by the speakers of that language. On the contrary, from a computational system point of view, the ability to deal with the variability of language expressions is not an easy task. To some extent, we could say that inferences related to linguistic phenomena could be added to the argument as new evidence to support the reasoning process (i.e. additional premises, expressing for instance that "house" and "habitation" are synonyms, or that the active/passive structures of a verb x are equivalent). Complex premises could therefore be decomposed into simpler premises, introducing the linguistic knowledge and the relations among premises and conclusions needed by a system to perform the inference task (Chapter 4 will discuss this issue in more detail).

2.9 Conclusion

In the light of the definitions provided in logic, the term "Textual Entailment" used in Computational Linguistics turns out to be somehow troublesome. Actually, TE involves both deductive and inductive arguments, the latter prevailing numerically on the first ones. Furthermore, also the motivation underlying the proposal of a generic framework to model language variability has been source of misunderstandings, since the definition of TE does not set a clear distinction line between linguistic knowledge and world knowledge that is involved in such kind of reasoning. In the Recognizing

Textual Entailment challenge (discussed in the next Chapter) strategies to deal with this issue have been outlined, partially guided by reasons of convenience for the task definition.

The four criteria for argument evaluation that we have applied to TE pairs have highlighted that *i)* in TE the premises are assumed to be true; *ii)* relevance is an essential criterion, even if simplifying assumptions have been made (i.e. same meaning of entities mentioned in T and H); *iii)* the criterion of total evidence sends back to the problem of background knowledge, since incomplete arguments require to be supported by new evidence both to validate or invalidate the conclusion.

The study of the types of arguments in logic allowed us to compare TE to categories of arguments that up to now have not been part of the research agenda (i.e. inductive arguments by analogy). Even if we will not discuss these aspects in the present Thesis, we have highlighted interesting perspectives for future work in this direction.

Finally, we pointed out that complex Ts can be usefully decomposed in simple premises. This process has the goal to highlight the relations among premises and conclusions, that are necessary from a computational system viewpoint. "Decomposing" will be somehow the leitmotif of the present Thesis and of the proposed "component-based" approach to TE.

Chapter 3

Recognizing Textual Entailment

This Chapter presents the state of the art of the research in Textual Entailment. Given the significant number of publications on this topic, we focus on the aspects of the previous works that are more relevant to the component-based framework for TE we propose in this Thesis.

3.1 Introduction

At the present time, textual entailment can be considered a hot topic within the Natural Language Processing community, as it represents an important field of investigation. High interest is demonstrated by:

- the Recognizing Textual Entailment (RTE) evaluation campaign, repeated yearly since 2005 (described in more details in Section 3.2);
- several publications on this topic, among the others Androutsopoulos and Malakasiotis (2010) [3], and Dagan *et al.* (2009) [27] provide an overview of the research in TE;
- a special issue of the *Journal of Natural Language Engineering*[1] on Textual Entailment (Volume 15, Special Issue 04) in 2009;

[1] http://journals.cambridge.org/action/displayIssue?jid=NLE&volumeId=15&seriesId=0&issueId=04

- the organization of workshops, such as the Workshop on Applied Textual Inference (*TextInfer*) at its second edition in 2011;[2]

- the organization of tutorials, such as the Tutorial on Recognizing Textual Entailment[3] at NAACL 2010;

- concerning languages different from English, the second evaluation campaign of Natural Language Processing tools for Italian (*EVALITA 2009*), supported by the NLP working group of AI*IA, added TE recognition among its tasks.[4]

In the previous chapter (Section 2.8, Chapter 2) we defined the notion of TE (Dagan and Glickman 2004 [28]), and the applications aimed at natural language processing and understanding that can benefit from this scenario. In this Chapter we focus on the Recognizing Textual Entailment (RTE) initiative, i.e. the evaluation framework for TE[5], and we provide an overview of the relevant work in the field (Section 3.2). In particular, we will focus on the works in the TE literature whose subject is more related to the content of the Thesis, i.e. previous analysis and annotations of the phenomena relevant to inference (Section 3.3).

3.2 The RTE Evaluation Campaign

In 2005, the PASCAL Network of Excellence started an attempt to promote a generic evaluation framework covering semantic-oriented inferences needed for practical applications, launching the Recognizing Textual Entailment (RTE) Challenge (Dagan *et al.* 2005 [29], Dagan *et al.* 2006

[2] http://sites.google.com/site/textinfer2011/

[3] http://naaclhlt2010.isi.edu/tutorials/t8.html

[4] http://evalita.fbk.eu/te.html

[5] For further information, see the Textual Entailment Resource Pool: http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

[30], Dagan *et al.* 2009 [27]) with the aim of setting a benchmark for the development and evaluation of methods that typically address the same type of problems but in different, application-oriented manners. As many of the needs of several Natural Language Processing applications can be cast in terms of TE (as discussed in Chapter 2), the goal of the evaluation campaign is to promote the development of general entailment recognition engines, designed to provide generic modules across applications. Since 2005, such initiative has been yearly repeated: **RTE-1** in 2005 (Dagan *et al.* 2005 [29]), **RTE-2** in 2006 (Bar-Haim *et al.* 2006 [6]) and **RTE-3** in 2007 (Giampiccolo *et al.* 2007 [40]), **RTE-4** in 2008 (Giampiccolo *et al.* 2008 [39])[6], **RTE-5** in 2009 (Bentivogli *et al.* 2009 [14])[7], and **RTE-6** in 2010 (Bentivogli *et al.* 2010 [12]).[8] Since 2008, RTE has been proposed as a track at the Text Analysis Conference (TAC)[9], jointly organized by the National Institute of Standards and Technology[10] and CELCT[11].

In this frame, which has taken a more explorative than competitive turn, the RTE task consists of developing a system that, given two text fragments (the *text* T and the *hypothesis* H), can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other. Example 3.1 represents a positive example pair, where the entailment relation holds between T and H (pair 10, RTE-4 test set). For pairs where the entailment relation does not hold between T and H, systems are required to make a further distinction between pairs where the entailment does not hold because the content of H is contradicted by the content of T (e.g. Example 3.2 - pair 6, RTE-4 test set), and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the

[6] `http://www.pascal-network.org/Challenges/\{RTE,RTE2,RTE3,RTE4\}`
[7] `http://www.nist.gov/tac/2009/RTE/`
[8] `http://www.nist.gov/tac/2010/RTE/`
[9] http://www.nist.gov/tac/about/index.html
[10] `http://www.nist.gov/index.html`
[11] `http://www.celct.it/`

content of T (e.g. Example 3.3 - pair 699, RTE-4 test set).

(3.1) T: *In the end, defeated, Anthony committed suicide and so did Cleopatra, according to legend, by putting an asp to her breast.*

H: *Cleopatra committed suicide.* **ENTAILMENT**

(3.2) T: *Reports from other developed nations were corroborating these findings. Europe, New Zealand and Australia were also beginning to report decreases in new HIV cases.*

H: *AIDS victims increase in Europe.* **CONTRADICTION**

(3.3) T: *Proposals to extend the Dubai Metro to neighbouring Ajman are currently being discussed. The plans, still in the early stages, would be welcome news for investors who own properties in Ajman.*

H: *Dubai Metro will be expanded.* **UNKNOWN**

This three-way judgement task (*entailment* vs *contradiction* vs *unknown*) was introduced since RTE-4, while before a two-way decision task (*entailment* vs *no entailment*) was asked to participating systems. However, the classic two-way task is offered as an alternative also in recent editions of the evaluation campaign (*contradiction* and *unknown* judgements are collapsed into the judgement *no entailment*). The submitted systems are tested against manually annotated data sets, which include typical examples that correspond to success and failure cases of NLP applications. In the data sets, the distribution according to the three way annotation is 50% entailment pairs, 35% unknown pairs, and 15% contradiction pairs (more details are provided in Section 3.2.1).

From year to year, the submissions have been numerous and diverse, as showed in Figure 3.1 that reports the number of participating systems.[12]

[12]In RTE-6, the main task is different from the previous ones. The number of participating teams is included in the graph, but the task is not comparable.

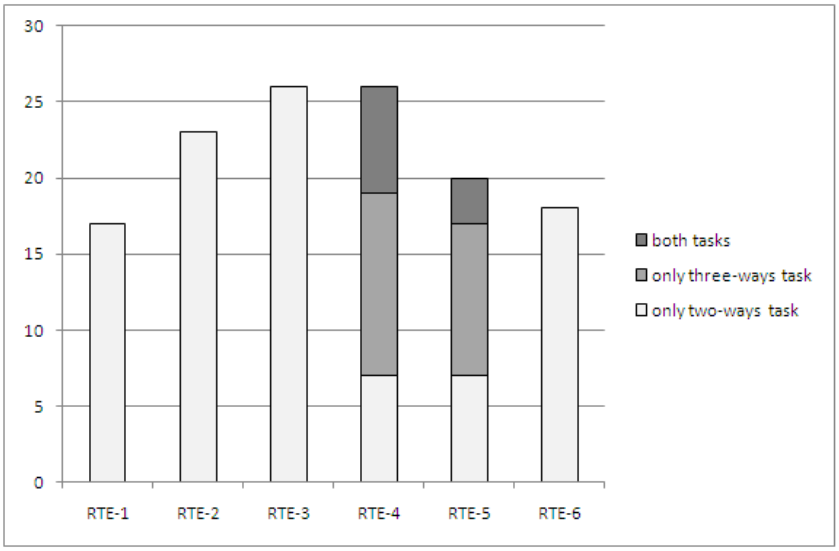


Figure 3.1: Systems participating in previous RTE challenges (main task)

TE systems are evaluated basing on their accuracy and, optionally, average precision, as a measure for ranking the pairs according to their entailment confidence. Figures 3.2 and 3.3[13] compare systems' results, respectively for two-ways and for three-ways judgement tasks, in the past editions of RTE[14], while Figure 3.4 shows the Word Overlap baseline for each data set[15] (Mehdad and Magnini 2009 [63]). As can be seen, on average systems's performances range from 55% to 65% of accuracy (not far from the baseline), meaning that current approaches are generally too simplistic with respect to the complexity of the task, and that there is still much room for improvement. General improvements with time can be noticed especially in first three editions. Then, stable performances of systems in RTE-4 and 5 are due to the introduction of longer and un-edited texts in the data sets, to make the task more challenging.

Beside the main task, that maintained the basic structure throughout

[13]Credits to RTE organizers (`http://www.nist.gov/tac/publications/2009/agenda.html`).

[14]RTE-6 is not considered, since the main task is different from the previous ones, and therefore not comparable. We will discuss about that lately in this Section.

[15]Calculated as H-T tokens, no stopwords, no normalization.

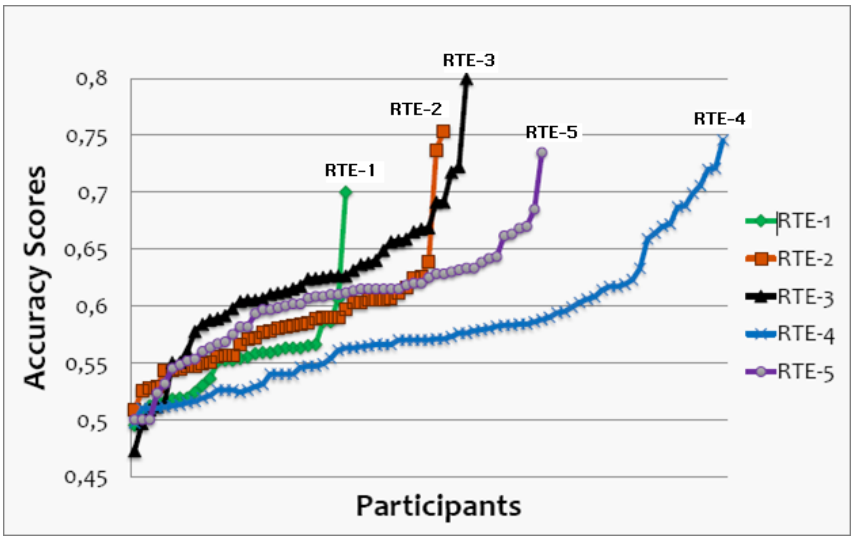


Figure 3.2: Systems' performances for the two-way judgement task

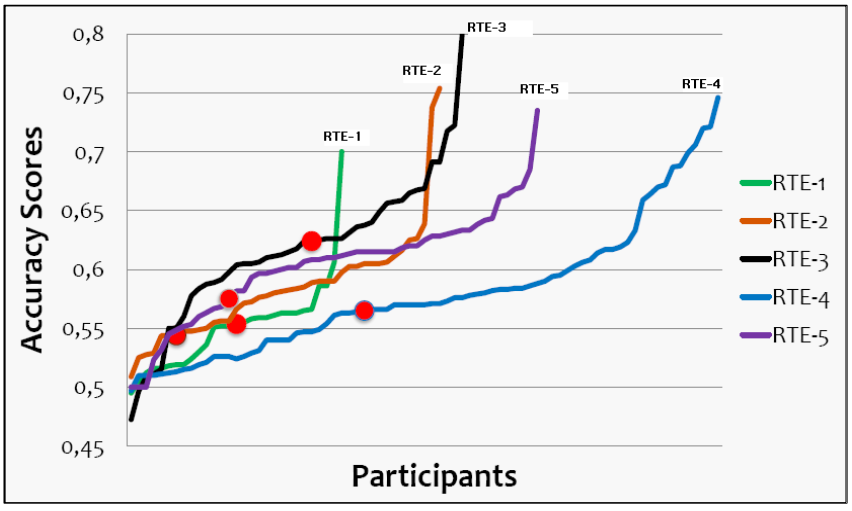


Figure 3.3: Baseline for the two-way judgement task

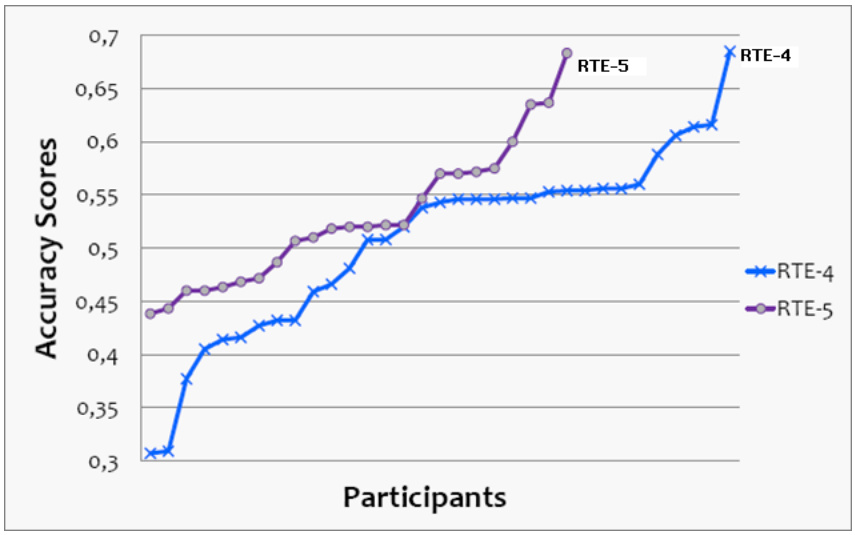


Figure 3.4: Systems' performances for the three-way judgement task

the editions of the challenge (except in RTE-6), a pilot task has been proposed from RTE-3 on (except in RTE-4), to experiment more realistic scenarios. RTE-3 Pilot task, called "Extending the Evaluation of Inference Texts", required the participating systems *i)* to give a more detailed judgement (i.e. three-way judgement task) against the same test set used in the main task, and *ii)* to provide justifications for the decisions taken. At RTE-5, a TE "Search Pilot task" was proposed, that consists in finding all the sentences that entail a given H in a given set of documents about a topic (i.e. the corpus). This task is situated in the summarization application setting, where *i)* H's are based on Summary Content Units (Nenkova *et al.* 2007 [73]) created from human-authored summaries for a corpus of documents about a common topic, and *ii)* the entailing sentences (T's), are to be retrieved in the same corpus for which the summaries were made.

In the following edition of the challenge, i.e. RTE-6, the Search Pilot task replaced the traditional main task. A new Pilot task was proposed at RTE-6, called "Knowledge Base Population Validation Pilot Task". It is situated in the Knowledge Base Population Scenario and aims to validate

the output of the systems participating in the KBP Slot Filling Task by using Textual Entailment techniques. In other words, systems are asked to determine whether a candidate slot filler is supported in the associated document using TE. With respect to the traditional setting, the pilot tasks impose new challenges to RTE systems developers, to make a step forward and to start to test RTE systems against real data.

In the next Sections, we describe in more details the traditional main task, focusing in particular on the data sets provided by the organizers of the challenge (Section 3.2.1), the approaches experimented by the participating teams (Section 3.2.2), linguistic/knowledge resources integrated in the systems (Section 3.2.3) and the tools used to pre-process the data (Section 3.2.4).

3.2.1 RTE data sets

The rationale underlying RTE data sets is that recognizing textual entailment should capture the underlying semantic inferences needed in many application settings (Dagan *et al.* 2009 [27]). For this reason, T-H pairs are collected from several applicative scenarios (e.g. Question Answering, Information Extraction, Information Retrieval, Summarization), reflecting the way by which the corresponding application could take advantage of automated entailment judgement. In the collection phase, each pair of the data set is judged by three annotators, and pairs on which the annotators disagree are discarded. On average, the final training and test data sets contain about 1000 pairs each, and the distribution according to the three-way annotation, both in the individual setting and in the overall data sets, is: 50% *entailment*, 35% *unknown*, and 15% *contradiction* pairs.

As discussed in Section 2.8.2, the definition of entailment in RTE pairs considers if a competent speaker with basic knowledge of the world would typically infer H from T. Entailments are therefore dependent on linguistic

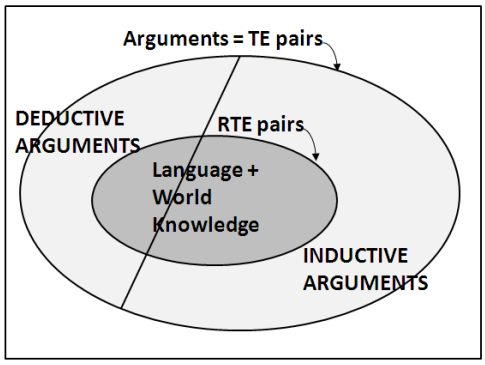


Figure 3.5: RTE data sets with respect to the distribution of logical arguments

knowledge, and may also depend on some world knowledge. Figure 3.5 represents the RTE data sets with respect to the arguments as defined in classical logic (Chapter 2) (see the controversy between Zaenen *et al.* 2005 [98] and Manning 2006 [58]). Partially guided by reasons of convenience for the task definition, some assumptions have been defined by the organizer of the challenge, as for instance, the a priori truth of the texts, and the same meaning of entities mentioned in T and H.

From a human perspective, the inference required are fairly superficial, since generally no long chains of reasoning are involved. However some pairs are designed to trick simplistic approaches (e.g. Bag of Words approaches), as showed in Example 3.4 (pair 397, RTE-2 test set).

(3.4) T: *Most of the open tombs in the Valley of the Kings are located in the East Valley, and this is where most tourists can be found.*

H: *The Valley of the Kings is located in the East Valley.*

Since the goal of RTE data sets is to collect inferences needed by NLP applications while processing real data, the example pairs are very different from a previous resource built to address natural language inference problems, i.e. the FraCas test suite (Cooper *et al.* 1996 [35]). This resource includes 346 problems, containing each one or more premises and one ques-

tion (i.e. the goal of each problem is expressed as a question).[16] With respect to RTE pairs, here the problems are designed to cover a broader range of semantic and inferential phenomena, including quantifiers, plurals, anaphora, ellipsis and so on, as showed in Example 3.5 (fracas-022: monotonicity, upwards on second argument).

(3.5) P1: *No delegate finished the report on time.*

Q: *Did no delegate finish the report?*

H: *No delegate finished the report.*

Answer: *unknown*

Why: *can't drop adjunct in negative context*

However, even if the FraCas test suite is much smaller when compared to the number of annotated pairs in RTE data sets, and it is less natural-seeming (i.e. it provides textbook examples of semantic phenomena, quite different from the kind of inferences that can be found in real data), it is worth mentioning it in this context.

3.2.2 RTE Approaches

A number of data-driven approaches applied to semantics have been experimented throughout the years, since the launch of the RTE Challenge in 2005. In general, the approaches still more used by the submitted systems include Machine Learning (typically SVM), logical inference, cross-pair similarity measures between T and H, and word alignment.

Machine Learning approaches (e.g. Kozareva and Montoya 2006 [50], Zanzotto *et al.* 2007 [100], Zanzotto *et al.* 2009 [101]) take advantage of the availability of the RTE data sets for training, and formulate TE as a classification task. A variety of features, including lexical-syntactic

[16] Bill MacCartney (Stanford University) converted FraCas questions into a declarative hypothesis: `http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`

and semantic features, are therefore extracted from training examples, and then used to build a classifier to apply to the test set for pair classification.

Other TE approaches underpin a transformation-based model, meaning that systems attempt to provide a number of transformations that allow to derive H from T. Different transformation-based techniques over syntactic representations of T and H have been proposed: for instance, (Kouylekov and Magnini 2005 [47]) assume 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), which are necessary to transform T into H. BarHaim *et al.* (2008) [5] model semantic inference as application of entailment rules in a transformation-based framework. Such rules, that specify the generation of entailed sentences from a source sentence, capture semantic knowledge about linguistic phenomena. Also (Harmeling *et al.* 2009 [44]) introduce a system for textual entailment that is based on a probabilistic model of entailment. This model is defined using a calculus of transformations on dependency trees, where derivations in that calculus preserve the truth only with a certain probability.

Another successful line of research to address TE is based on deep analysis and semantic inference. Different approaches can be considered part of this group: *i)* approaches based on logical inferences (e.g. Tatu and Moldovan 2007 [88], Bos and Markert 2006 [16]); *ii)* application of natural logic (e.g. Chambers *et al.* 2007 [20], MacCartney 2009 [53]); *iii)* approaches exploiting ontology-based reasoning (e.g. Sibilini and Kosseim 2008 [85]). Such approaches are generally coupled with data-driven techniques, where the final decision about the entailment relation is taken on the basis of semantic features managed by Machine Learning algorithms.

Some experimented approaches to RTE use vector space model of semantics, meaning that each word of the input pairs is mapped to a vector,

that shows how strongly the words co-occur with particular other words in the corpora (Lin 1998 [52]). Syntactic information can be considered: for example, in (Padó and Lapata 2007 [77]) co-occurring words are required to participate in particular syntactic dependencies. A compositional vector-based meaning representation theory can then be used to combine the vector of single words (Mitchell and Lapata 2008 [69]).

Similar in spirit to the research direction we propose in this Thesis, a component-based system has been developed by (Wang and Neuman 2008 [94]), based on three specialized RTE-modules: *i)* to tackle temporal expressions; *ii)* to deal with other types of NEs; *iii)* to deal with cases with two arguments for each event. Besides these precision-oriented modules, two robust but less accurate backup strategies are considered, to deal with not yet covered cases. In the final stage, the results of all specialized and backup modules are joint together, applying a weighted voting mechanism.

3.2.3 Knowledge resources

Lexical databases, such as WordNet (Fellbaum 1998 [36])[17], EuroWordNet[18], and eXtended WordNet[19] are among the most used resources by TE systems. Also DIRT (Discovery of Inference Rules from Text) (Lin and Pantel 2001 [51])[20], a collection of inference rules, is used by several systems (e.g. Clark and Harrison 2008 [22], Mirkin *et al.* 2009 [65]), as well as verb-oriented resources such as VerbNet[21] (e.g. Balahur *et al.* 2008 [4]), and VerbOcean[22] (e.g. Wang *et al.* 2009 [96]). Also FrameNet[23] was integrated in some systems (e.g. Delmonte *et al.* 2007 [32]), although in a

[17] http://wordnet.princeton.edu/
[18] http://www.illc.uva.nl/EuroWordNet/
[19] http://xwn.hlt.utdallas.edu/
[20] http://www.aclweb.org/aclwiki/index.php?title=DIRT_Paraphrase_Collection
[21] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[22] http://demo.patrickpantel.com/demos/verbocean/
[23] http://framenet.icsi.berkeley.edu/

limited way probably because of its restricted coverage or of the difficulties in modelling FrameNet information (see also Burchardt *et al.* 2009 [17]). On the contrary, in the last editions of the Challenge, it was possible to notice an increasing tendency in considering the web as a resource. Many participating systems used information from Wikipedia to extract entailment rules, Named Entities and background knowledge (e.g. Bar-Haim *et al.* 2008 [5], Mehdad *et al.* 2009 [64]).

To better understand the kind of knowledge resources most frequently used by participating systems, and their contribution in recognizing textual entailment, the RTE Challenge organizers have created a dedicated website containing a repository of linguistic tools and resources for TE, i.e. the Textual Entailment Resource Pool.[24] Moreover, in order to evaluate the contribution of each single resource to the systems' performances, ablation tests were introduced as a requirement for systems participating in RTE-5 and RTE-6 main tasks. Ablation tests consist in removing one module at a time from a system, and re-running the system on the test set with the other modules, except the one tested. Unluckily, the results obtained from ablation tests are not straightforward in determining the actual impact of the resources, since the different uses made by the systems of the same resources, make it difficult to compare the results.

3.2.4 Tools for RTE data preprocessing

Various tools are generally used to pre-process the RTE pairs of the data sets, so their accuracy can have a strong impact on TE system performances. Among the most frequently used tools there are Part-of-Speech taggers such as TextPro[25] (e.g. Mehdad *et al.* 2009 [64]) and SVM tag-

[24] http://www.aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources
[25] http://textpro.fbk.eu/

ger[26] (e.g. Yatbaz 2008 [97]); parsers such as Minipar[27], Stanford Parser[28] (e.g. Wang *et al.* 2009 [96]); stemmer such as Porter's stemmer[29]; Named Entity Recognizer such as Stanford NER.[30] Also software tools such as WEKA[31] for Machine Learning approaches, and Lucene for indexing, are largely used (e.g. Bar-Haim *et al.* 2008 [5]), as well as WordNet similarity tools. A list of the tools mostly used by participating systems can be found in the Textual Entailment Resource Pool web page.[32]

3.3 Analysis of phenomena relevant to inference

As introduced before, the example pairs in RTE data sets represent different levels of entailment reasoning, such as lexical, syntactic, morphological and logical. Several studies in the literature have tried to analyse such linguistic levels in relation to the recognizing textual entailment task.

In Garoufi (2007) [38], a scheme for manual annotation of textual entailment data sets (ARTE) is proposed, with the aim of highlighting a wide variety of entailment phenomena in the data. ARTE views the entailment task in relation to three levels, i.e. *Alignment*, *Context* and *Coreference*, according to which 23 different features for positive entailment annotation are extracted. Each level is explored in depth for the positive entailment cases, while for the negative pairs a more basic and elementary scheme is conceived. The ARTE scheme has been applied to the complete positive entailment RTE-2 test set (400 pairs, i.e. 100 pair of each task), and to a random 25% portion of the negative entailment test set, equally distributed

[26] http://www.ling.upenn.edu/~beatrice/corpus-ling/svmt.html
[27] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm
[28] http://nlp.stanford.edu/software/lex-parser.shtml
[29] http://tartarus.org/~martin/PorterStemmer/
[30] http://nlp.stanford.edu/ner/index.shtml
[31] http://www.cs.waikato.ac.nz/ml/weka/
[32] http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool\#Tools

among the four tasks (100 pairs, i.e. 25 pairs of each task). *Reasoning* is the most frequent feature appearing altogether in 65.75% of the annotated pairs: this indicates that a significant portion of the data involves deeper inferences. The combination of the entailment features is analysed together with the entailment types and their distribution in the data.

An attempt to isolate the set of T-H pairs whose categorization can be accurately predicted based solely on syntactic cues has been carried out in (Vanderwende *et al.* 2005 [91]). Aim of this work is to understand what proportion of the entailments in the RTE-1 test set could be solved using a robust parser. Two human annotators evaluated each T-H pair of the test set, deciding whether the entailment was: *true by syntax*; *false by syntax*; *not syntax*; *can't decide*. Additionally, annotators were allowed to indicate whether the recourse to information in a general purpose thesaurus entry would allow a pair to be judged true or false. Their results show that 37% of the test items can be handled by syntax, broadly defined (including phenomena such as argument assignment, intra-sentential pronoun anaphora resolution); 49% of the test items can be handled by syntax plus a general purpose thesaurus. According to their annotators, it is easier to decide when syntax can be expected to return *true*, and it is uncertain when to assign *false*. Basing on their own observations, the submitted system (Vanderwende *et al.* 2006 [92]) predicts entailment using syntactic features and a general purpose thesaurus, in addition to an overall alignment score. The syntactic heuristics used for recognizing false entailment rely on the correct alignment of words and multiwords units between T and H logical forms.

Bar Haim *et al.* (2005) [8] define two intermediate models of TE, which correspond to lexical and lexical-syntactic levels of representation. Their lexical level captures knowledge about lexical-semantic and morphological relations, and lexical world knowledge. The lexical-syntactic level additionally captures syntactic relationships and transformations, lexical-syntactic

inference patterns (rules) and co-reference. They manually annotated a sample from the RTE-1 data set according to each model, compared the outcomes for the two models as well as for their individual components, and explored how well they approximate the notion of entailment. It was shown that the lexical-syntactic model outperforms the lexical one, mainly because of a much lower rate of false-positives, but both models fail to achieve high recall. The analysis also showed that lexical-syntactic inference patterns stand out as a dominant contributor to the entailment task.

Also (Clark *et al.* 2007 [24]) agree that only a few entailments can be recognized using simple syntactic matching, and that the majority rely on significant amount of the so called "common human understanding" of lexical and world knowledge. The authors present an analysis of 100 (25%) of the RTE-3 positive entailment pairs, to identify where and what kinds of world knowledge are needed to fully identify and justify the entailment, and discuss several existing resources (see Section 3.2.3) and their capacity for supplying that knowledge. After showing the frequency of the different entailment phenomena from the sample they analysed, they state that very few entailments depend purely on syntactic manipulation and a simple lexical knowledge (synonyms, hypernyms), and that the vast majority of entailments require significant world knowledge.

In (Dagan *et al.* 2008 [26]), where a framework for semantic inference at the lexical-syntactic level is presented, the authors show that the inference module can be exploited also for improving unsupervised acquisition of entailment rules through canonization (i.e. the transformation of lexical-syntactic template variations that occur in a text into their canonical form - this form is chosen to be the active verb form with direct modifier). The canonization rule collection is composed by two kinds of rules: *i)* syntactic-based rules (e.g. passive/active forms, removal of conjunctions, removal of appositions), *ii)* nominalization rules, trying to capture the relations

between verbs and their nominalizations. The authors propose to solve the learning problems using this entailment module at learning time as well.

A definition of contradiction for TE task is provided by (de Marneffe *et al.* 2008 [59]), together with a collection of contradiction corpora. Detecting contradiction appears to be a harder task than detecting entailment, since it requires deeper inferences, assessing event coreference and model building. Contradiction is said to occur when two sentences are extremely unlikely to be true simultaneously; furthermore, they must involve the same event. A previous work on the same topic was presented by (Harabagiu *et al.* 2006 [43]), in which the first empirical results for contradiction detection were provided (they focused only on specific kind of contradiction, i.e. those featuring negation and those formed by paraphrases).

Kirk (2009) [45] describes his work of building an inference corpus for spatial inference about motion, while Wang and Zhang (2008) [95] focus on recognizing TE involving temporal expressions. Akhmatova and Dras (2009) [2] experiment current approaches on hypernymy acquisition to improve entailment classification.

Basing on the intuition that frame-semantic information is a useful resource for modelling textual entailment, (Buchardt *et al.* 2009 [17]) provide a manual frame-semantic annotation for the test set used in RTE-2 (i.e. the FATE corpus) and discuss experiments conducted on this basis.

Bentivogli *et al.* (2009) [13] focus on some problematic issues related to resolving coreferences to entities, space, time and events at the corpus level, as emerged during the annotation of the data set for the textual entailment Search Pilot. Again at the discourse level, (Mirkin *et al.* 2010 [67], and Mirkin *et al.* 2010b [66]) analyse various discourse references in entailment inference (manual analysis on RTE-5 data set) and show that while the majority of them are nominal coreference relations, another substantial part is made up by verbal terms and bridging relations.

3.3.1 Sammons *et al.* 2010 [83]

Researchers at the University of Illinois recently carried out an annotation work that is very similar in spirit to the approach we propose in this Thesis (that will be described in details in Chapter 6). Highlighting the need of resources for solving textual inference problems in the context of RTE, Sammons *et al.* 2010 [83] challenge the NLP community to contribute to a joint, long term effort in this direction, making progress both in the analysis of relevant linguistic phenomena and their interaction, and developing resources and approaches that allow more detailed assessment of RTE systems. The authors propose a linguistically-motivated analysis of entailment data based on a step-wise procedure to resolve entailment decision, by first identifying parts of T that match parts of H, and then identifying connecting structures. Their inherent assumption is that the meanings of T and H could be represented as sets of n-ary relations, where relations could be connected to other relations (i.e. could take other relations as arguments). The authors carried out a feasibility study applying the procedure to 210 examples from RTE-5, marking for each example the entailment phenomena that are required for the inference.[33]

3.4 Conclusions

In this Chapter we presented the state of the art of the research in Textual Entailment, providing some pointers to stress the current interest of the research community on this topic. In particular, we described the Recognizing Textual Entailment challenge, that since 2005 represents the evaluation framework for TE systems. Although several approaches have been experimented, and tools and resources have been developed to provide more knowledge to solve the inference task, systems performances are still

[33] https://agora.cs.illinois.edu/display/rtedata/Annotation+Resources

far from being optimal (the accuracy of most of them ranges between 55% to 65% for the two-way judgement task). While on one side the tasks proposed by the organizers of the challenge are of increasing difficulty to move towards more real scenarios, on the other side TE systems capabilities are not improving accordingly. For this reason, a renewed interest is rising in the TE community towards a more fine-grained analysis of the phenomena underlying the entailment/contradiction relations, and the goal of the next Chapters of this Thesis is to analyse and provide some contributions on different dimensions of the problem.

Chapter 4

A Component-Based Framework for Textual Entailment

In this Chapter we propose a framework for component-based Textual Entailment, and we show that decomposing the complexity of TE focusing on single phenomena involved in the inference relation, and on their combination, brings interesting elements to advance in the comprehension of the main task.

4.1 Introduction

In Chapter 3 we discussed the main approaches that have been experimented to face the RTE task, and we highlighted the progresses in TE technologies that have been shown in past RTE evaluation campaigns. Nevertheless, a renewed interest is rising in the TE community towards a deeper and better understanding of the core phenomena involved in textual inference. In line with this direction, we are convinced that crucial progress may derive from a focus on decomposing the complexity of the TE task into basic phenomena and on their combination. This belief demonstrated to be shared by the RTE community, and a number of recently published works (e.g. Sammons *et al.* 2010 [83]) agree that incremental advances

in local entailment phenomena are needed to increase the performances in the main task, which is perceived as omni-comprehensive and not fully understood yet.

The intuition underlying the *component-based framework* for TE we propose, is that the more a system is able to correctly solve the linguistic phenomena relevant to the entailment relation separately, the more the system should be able to correctly judge more complex pairs, in which different phenomena are present and interact in a complex way. Such intuition is motivated by the notion of meaning compositionality, according to which the meaning of a complex expression is determined by its structure and by the meaning of its constituents (Frege 1992 [37]). In a parallel way, we assume that it is possible to recognize the entailment relation of a T-H pair (i.e. to correctly judge the *entailment/contradiction* relation) only if all the phenomena contributing to such a relation are resolved. Analysing once again the TE pairs in the light of our study on logical arguments, we show how complex Ts can be usefully decomposed into simple premises, that can be added to the argument to provide either the world knowledge or the linguistic evidence needed by a computational system to infer the conclusion through intermediate inferential steps (Section 4.2). The interactions and the dependencies among the linguistic phenomena in a pair are considered while combining the partial steps to obtain the final judgement for a pair (Section 4.3).

In Section 4.4 we define a general architecture for component-based TE, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. Although no specific constraints are defined with respect to how such components should be implemented, our proposal focuses on a transformation-based approach, that we define taking advantage of the conceptual and formal tools available from an extended model of Natural Logic (NL) (MacCartney and

Manning 2009 [56]) (Section 4.5). Given a T-H pair, each TE component performs *atomic edits* to solve the specific linguistic phenomenon it is built to deal with, and assigns an entailment relation as the output of this operation. We provide an operational definition of atomic edits allowed for a specific phenomenon in terms of application of entailment rules. Once the TE components have assigned an entailment relation to each phenomena relevant to inference in a specific pair, NL mechanisms of semantic relations composition are applied to join the output of each single component, in order to obtain the final entailment judgement for a pair.

4.2 Decomposing the TE task

In Chapter 2, the study of the types of arguments in logic allowed us to compare TE pairs to certain categories of arguments, and to evaluate them according to the criteria described in Nolt *et al.* (1998) [75]. Taking advantage of those observations and definitions, in this section we motivate our proposal of decomposing complex TE pairs into simple premises, each conveying the world knowledge or the linguistic evidence required by a system to derive the conclusion through a chain of reasoning steps.

4.2.1 Towards total evidence: atomic arguments

Most arguments in natural language discourse are incompletely expressed, i.e. they can be thought of as having unstated assumptions (Nolt *et al.* 1998 [75]). Missing premises or conclusions that are assumed by the argument are intended to be so obvious as to not need stating. In other words, the speaker avoids alienating listeners with long chains of inferences and appeal to the audience's common sense without reducing the logical force of

the argument (Walton and Reed 2005 [93]).[1] Many examples of arguments with missing premises are in fact based on assumptions that come under the heading of common knowledge, i.e. everyday human experience of the way things generally work, about familiar human intuitions and values, and about the way we can expect most people to generally react. While humans can easily cope with most cases of argument incompleteness, for an automatic system this is anything but an easy task.

A strategy to add missing premises in incomplete arguments expressed in natural language should therefore be thought, in order to fill the gap between the given premises and the conclusion to be proved. To support the reasoning process of automatic systems, also evidences at a fine-grained level should be provided, meaning that both the linguistic and the world knowledge required to infer the conclusion should be made explicit and added as premises. To some extent, for computational purposes we need to take the requirement of total evidence - Criterion 4, discussed in Chapter 2 - to extremes.

While remaining faithful to what we know of the arguer's thought, i.e. the content of T and H in TE pairs expressed as logical arguments, we try to make the argument as strong as possible following the principle of charity (Chapter 2). We propose *i)* to simplify complex Ts through decomposition, and *ii)* to fill in the missing premises that provide the pieces of evidence needed by a system to infer the conclusion through a chain of inferential steps. Implicit premises concerning both the linguistic and the world knowledge required by the inference task in a specific argument are therefore made explicit and added to the argument. Such premises should allow a system to carry out a step of reasoning on a particular sub-problem of entailment, and to derive a conclusion. This conclusion

[1] In particular, this paper explores the role of argumentation schemes in the so-called *enthymeme* (i.e. arguments with missing premises or conclusions) reconstruction.

can then function as a premise for yet another conclusion, and so on, as in complex arguments (described in Chapter 2). More precisely, starting from the original argument, a complex premise is decomposed into a set of simpler premises (*nonbasic premises* or *intermediate conclusions*), each allowing to carry out an inferential step on a sub-portion of the original premise focusing on a specific phenomenon relevant to derive the conclusion. At each step the piece of knowledge or of linguistic evidence needed to correctly infer the (intermediate) conclusion is made explicit and added to the argument as new premise. The final conclusion is therefore inferred through a chain of simple steps of reasoning from the given premises along with the missing premises. Each of the simple steps of reasoning, which are linked together to form a complex argument, is an argument in its own right. Since they express the minimal inferential step related to a sub-problem of entailment, we define them *atomic arguments* (AA). To be considered atomic, an argument should require only the minimal piece of knowledge (added as new premise) needed to derive the conclusion from the original premise. The structure of an atomic argument can be schematized as follows:

AA
(1) PREMISE
(2) ADDITIONAL PREMISE (IMPLICIT ASSUMPTION)
∴ (3) CONCLUSION

If more pieces of evidence should be provided to infer the conclusion, the argument is not atomic and it should be further decomposed. The process of decomposition of complex arguments into atomic arguments ends when no further decomposition of the original premise is possible, and when no more pieces of evidence (i.e. additional premises) are needed to derive the conclusion.

Premises providing new evidence on linguistic and world knowledge can be added provided that they are *true* and *pertinent*, i.e. that they are

compliant with Criteria 1 and 3 (described in Chapter 2).[2] The following scheme represents the structure of a complex argument, A, once decomposed into atomic arguments:

A [
- AA_1 [
  - (1) ORIGINAL PREMISE
  - (2) ADDITIONAL PREMISE (IMPLICIT ASSUMPTION)
  - $\therefore$ (3) INTERMEDIATE CONCLUSION (NON-BASIC PREMISE)
- AA_2 [
  - (4) ADDITIONAL PREMISE (IMPLICIT ASSUMPTION)
  - $\therefore$ (j) INTERMEDIATE CONCLUSION (NON-BASIC PREMISE)
- AA_n [
  - (k) ADDITIONAL PREMISE (IMPLICIT ASSUMPTION)
  - $\therefore$ (m) FINAL CONCLUSION

Since each atomic argument is an argument in its own right (e.g. AA_1, AA_2, AA_n), it can be either deductive or inductive, according to Criterion 2. The properties of the initial argument should be maintained through the inference chain, so that the reasoning through intermediate conclusions is made easier, but not distorted.

Since we showed that we can consider TE pairs in the same way as arguments, we apply the same strategy with the goal of highlighting the relations between T and H through decomposition. Let's consider Example 4.1 (pair 408, RTE-5 test set [14]):

(4.1) T: *British writer Doris Lessing, recipient of the 2007 Nobel Prize in Literature, has said in an interview that the terrorist attack on September 11 "wasn't that terrible" when compared to attacks the Irish Republican Army (IRA) made on Britain [...].*

H: *Doris Lessing won the Nobel Prize in Literature in 2007.*

[2] Walton and Reed (2005) [93] discuss about the validity of incomplete arguments once the missing parts are filled in, and about the truth of the missing premises. The authors claim that from a pragmatic viewpoint, incomplete arguments should be filled in with missing assumptions that are *i)* plausible to the intended audience or recipient of the argument, and *ii)* that appear to fit in with the position advocated by the arguer, as far as the evidence of the text indicates. It is possible that the most natural candidate for the missing premise in an argument is a statement that it is false, or at least highly questionable: in this case the argument can come out as a bad one once completed.

we can represent it into the argument standard format as:

(4.2) *British writer Doris Lessing, recipient of the 2007 Nobel Prize in Literature, has said in an interview [...]*[3]

∴ *Doris Lessing won the Nobel Prize in Literature in 2007.*

According to our proposal, we should identify the missing pieces of linguistic and world knowledge evidence in the pair that are relevant to correctly derive the conclusion. At a fine-grained level, to be able to infer H from T in Example 4.1 we need to provide knowledge related to the different way it is possible to express a syntactic realization (T: *2007 Nobel Prize in Literature* ⇒ H: *Nobel Prize in Literature in 2007*). Furthermore, knowledge related to the syntactic phenomenon of apposition (T: *Doris Lessing, recipient of* ⇒ H: *Doris Lessing is the recipient of*) should be provided and solved through an intermediate inferential step. On the bases of this outcome (that we call T') other linguistic pieces of evidence concerning the verbalization process should be provided to carry out another step (T': *Doris Lessing is the recipient of* ⇒ H: *Doris Lessing received*). Again, the new outcome (that becomes T") should be used for the last step, where pieces of evidence concerning the general inference "x receive a prize" and "x won a prize" should be added in order to correctly state that H follows from T (T": *Doris Lessing received* ⇒ H: *Doris Lessing won*). These passages can be represented into the argument standard format, as:

(4.3) (1) *British writer Doris Lessing, recipient of the 2007 Nobel Prize in Literature, has said in an interview [...].*

(2) *2007 Nobel Prize in Literature* express the same meaning as *Nobel Prize in Literature in 2007*

∴ (3) *British writer Doris Lessing, recipient of the Nobel Prize in Literature*

[3]Often entailment pairs are interspersed with material extraneous to the argument. In such cases, we report only the relevant part.

in 2007 [...].

(4) *Doris Lessing, recipient of* express the same meaning as *Doris Lessing is the recipient of*

∴ (5) *British writer Doris Lessing is the recipient of the Nobel Prize in Literature in 2007.*

(6) *Doris Lessing is the recipient of* express the same meaning as *Doris Lessing received*

∴ (7) *British writer Doris Lessing received the Nobel Prize in Literature in 2007.*

(8) *Doris Lessing received* express the same meaning as *Doris Lessing won*

∴ (9) *Doris Lessing won the Nobel Prize in Literature in 2007.*

Statement (1) is the original T, and statements (2),(4),(6),(8) are the implicit premises we made explicit to provide the linguistic knowledge needed for computational purposes. An intermediate conclusion, i.e. (3),(5),(7), follows from each of these premises, meaning that an intermediate inferential step is carried out. Through intermediate steps we decompose the complexity of the task to derive the original conclusion (9).

It is possible that the starting argument is not a valid one, for different reasons discussed in Chapter 2 (i.e. either one of the premise contradicts the conclusion, or the inductive probability is too low to support the conclusion, or the conclusion is not pertinent). While decomposing the original argument according to our proposal, it can therefore be the case that one (or more) atomic arguments are not valid, breaking the reasoning chain. This can happen either when the additional premise provide linguistic or world knowledge evidence that invalidate the conclusion, or when we are not able to provide enough evidence to support the conclusion with a high inductive probability (i.e. for instance, if the conclusion contains more specific information with respect to the original premise).

4.2.2 Linguistic phenomena relevant to inference

Atomic arguments are characterised by a simple additional premise expressing the piece of linguistic or world knowledge evidence needed to derive the conclusion from the original premise. A categorization of these pieces of evidence is therefore crucial to allow, by translation, for a classification of the atomic arguments theirselves.

To have a clearer idea of the typology of the missing pieces of evidence that are required to infer the conclusion (H) from the premise (T) in TE pairs, we randomly extracted a sample of RTE pairs (30 entailment pairs, 30 contradiction and 30 unknown pairs) from RTE-5 test set (Bentivogli *et al.* 2009 [14]), and we decomposed them as explained in Section 4.2. For computational purposes we need a refined analysis of the missing evidence, that focuses mainly on the linguistic phenomena and the world knowledge required to support the reasoning process. Although different levels of granularity can be used to define the inference sub-problems, in this Thesis we decided to group the phenomena using both fine-grained categories and broader categories (Bentivogli *et al.* 2010 [11]). Macro categories are defined referring to widely accepted linguistic categories in the literature (e.g. Garoufi 2007 [38]) and to the inference types typically addressed in RTE systems: lexical, syntactic, lexical-syntactic, discourse and reasoning. Each macro category includes fine-grained phenomena, which are listed below. This list is not exhaustive and reflects the phenomena we detected in the sample of RTE-5 pairs we analysed.

- *lexical:* identity, format, acronymy, demonymy, synonymy, semantic opposition, hyperonymy, geographical knowledge;
- *lexical-syntactic:* nominalization/verbalization, causative, paraphrase, transparent heads;
- *syntactic:* negation, modifier, argument realization, apposition, list,

coordination, active/passive alternation;

- *discourse:* coreference, apposition, zero anaphora, ellipsis, statements;
- *reasoning:* apposition, modifiers, genitive, relative clause, elliptic expressions, meronymy, metonymy, membership /representativeness, reasoning on quantities, temporal and spatial reasoning, all the general inferences using background knowledge.

Some phenomena (e.g. apposition) can be classified in more than one macro category, according to their specific occurrence in the text. For instance, in Example 4.4 (Pair 8, RTE-5 test set):

(4.4) T: *The government of Niger and Tuareg rebels of the Movement of Niger People for Justice (MNJ) have agreed to end hostilities [...].*

H: *MNJ is a group of rebels.*

the apposition is considered as syntactic, while in Example 4.5:

(4.5) T: *Ernesto, now a tropical storm, made landfall along the coastline of the state of North Carolina [...].*

H: *Ernesto is the name given to a tropical storm.*

the apposition is classified into the category reasoning.[4]

It is worthwhile to note that since world knowledge is an omni-pervasive phenomenon (as discussed in Section 2.8.2), it has not been categorized separately. In our framework, the phenomena categorized above define the atomic inferential steps (atomic arguments) in which complex arguments should be decomposed.

[4]More details on the analysis we carried out and on the distribution of each phenomenon in the sample are provided in Chapter 6.

4.2.3 Entailment rules

As discussed in the previous sections, we assume that we can introduce linguistic and world knowledge evidence to the argument in the form of additional premises, to provide the information required by a system to support the reasoning process. For computational purposes, such knowledge can be expressed through entailment rules (Szpektor *et al.* 2007 [86]). An entailment rule is either a directional or bidirectional relation between two sides of a pattern, corresponding to text fragments with variables (typically phrases or parse sub-trees, according to the granularity of the phenomenon they formalize). The left-hand side of the pattern (LHS) entails the rights-hand side (RHS) of the same pattern under the same variable instantiation. In addition, a rule may be defined by a set of constraints, representing variable typing (e.g. PoS, Named Entity type) and relations between variables, which have to be satisfied for the rule to be correctly applied. A rule can have an associated probability, expressing the degree of confidence that its application preserves the entailment relation between T and H (e.g. in a range from 0 to 1). For instance, the entailment rule for demonyms can be expressed as:

Entailment rule:	**demonymy**
Pattern:	*X Y* $\Leftrightarrow$ *X (is) from Z*
Constraint:	*DEMONYMY(X,Z)*
	TYPE(X)= ADJ_NATIONALITY
	TYPE(Z)=GEO
Probability:	1

meaning that X Y entails Y *is from* Z if there is a ENTAILMENT relation of demonymy between X and Y, X is an adjective expressing a nationality and Z is a geographical entity (e.g. *A team of European astronomers* $\Leftrightarrow$ *A team of astronomers from Europe*, pair 205 RTE-5). The probability that

the application of such rule preserves the entailment relation is equal to 1.

The entailment rules for a certain phenomenon aim to be as general as possible, but for the cases in which the semantics of the specific words is essential (e.g. general inference based on common background), text snippets extracted from the data are used. In our framework, the entailment rules provide the minimal piece of knowledge or of linguistic evidence needed to derive a conclusion from a premise in an atomic argument. Different rules can be needed to formalize the variants in which the same phenomenon occurs in the pairs. For example, both the following entailment rules formalize the phenomenon of apposition (syntax):

Entailment rule:	**apposition_1**
Pattern:	*X, Y ⇔ Y X*
Constraint:	*APPOSITION(Y,X)*
Probability:	1

Entailment rule:	**apposition_2**
Pattern:	*X, Y ⇔ Y is X*
Constraint:	*APPOSITION(Y,X)*
Probability:	1

A possible instantiation of rule a) is: *Girija Prasad Koirala, Prime Minister ⇔ Prime Minister Girija Prasad Koirala*, while a possible instantiation of rule b) is: *Kim Iong II, the leader of North Korea ⇔ The leader of North Korea is Kim Iong II.*

4.2.4 Contradiction rules

As discussed in Section 4.2.1, while decomposing the original argument according to our proposal, it can be the case that one (or more) resulting

atomic arguments are not valid. In the cases in which it happens because the conclusion contradicts the premise, the linguistic and world knowledge pieces of evidence that support the reasoning process are still required by a computational system, but this time they should provide information about the mismatching situation. In a specular way with respect to entailment rules, we can express such knowledge in the form of *contradiction rules*. In this case, the associated probability expresses the degree of confidence that the application of the rule generates a contradiction relation between T and H. For instance, the contradiction rule for antonymy (i.e. semantic opposition) can be expressed as:

Contradiction rule:	**antonymy**
Pattern:	$X \nleftrightarrow Y$
Constraint:	*ANTONYMY(X,Y)*
Probability:	1

and can be instantiated as *east of Bergen* $\nleftrightarrow$ *west of Bergen*.

Another reason for which the atomic arguments obtained through the decomposition process can be not valid is that the inductive probability is too low to support the conclusion. In this case, the piece of evidence expressed by the rule is not sufficient to support the conclusion, i.e. the degree of confidence that the application of the rule preserves the entailment relation between T and H is very low. Collecting such kind of rules with a low probability does not really make sense for computational purposes, since we can somehow obtain them in a complementary way with respect to high-probability rules. In other words, if a certain rule is not present among the highly probable ones, it means that it has a low probability, and therefore it is not strong enough to support the related inferential step. The resulting atomic argument cannot be considered a "good" one, according to the criteria described in Chapter 2.

4.2.5 Atomic RTE pairs

The linguistic knowledge expressed in the form of entailment rules should provide the pieces of evidence needed to carry out a step of reasoning on a particular sub-problem of entailment present in a certain T-H pair. The goal is to derive an intermediate conclusion where the entailment relation conveyed by the phenomenon under consideration is solved. As introduced before, each of the simple steps of reasoning is therefore an argument in its own right, where a certain phenomenon relevant to the inference task is highlighted and isolated (i.e. atomic argument). We are convinced that having the possibility to derive such atomic arguments for all the phenomena that play an important role in the inference task - deriving them from original RTE pairs - could bring several advantages to TE system developers, that could profitably use them to train and evaluate ad hoc modules able to deal with sub-problems of TE.

For this reason, we propose a methodology for the creation of atomic arguments, that in the context of textual entailment we call *atomic T-H pairs*, i.e. pairs in which a certain phenomenon relevant to the entailment relation is highlighted and isolated (Magnini and Cabrio 2009 [57], Bentivogli *et al.* 2010 [11]).[5] The procedure consists of a number of steps carried out manually. We start from a T-H pair taken from one of the RTE data sets and we decompose T-H in a number of atomic pairs T-H_i, where T is the original Text and H_i are Hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in T-H. The procedure is schematized in the following steps:

1. Individuate the linguistic phenomena which contribute to the entailment in T-H.

[5] In our previous papers, we used to refer to the atomic T-H pairs as *monothematic pairs*. In this Thesis we decided to switch the terminology to be compliant with the theorical framework we propose.

2. For each phenomenon i:

   (a) individuate a general entailment rule r_i, and instantiate the rule using the portion of T which expresses i as the Left Hand Side (LHS) of the rule, and information from H on i as the Right Hand Side (RHS) of the rule.

   (b) substitute the portion of T that matches the LHS of r_i with the RHS of r_i.

   (c) consider the result of the previous step as H_i, and compose the atomic pair $T - H_i$. Mark the pair with phenomenon i.

3. Assign an entailment judgement to each atomic pair.

For instance, the decomposition of the pair in Example 4.1 (pair 408 in RTE-5) into atomic pairs can be schematized as follows:[6]

AA$_1$ [
T *Doris Lessing, recipient of the 2007 Nobel Prize in Literature [...]*
synt:arg_realiz X Y ⇔ Y IN X, TYPE(X)=TEMPORAL_EXPRESSION
∴ H_1 Doris Lessing, recipient of the Nobel Prize in Literat. in 2007

AA$_2$ [
synt:apposition X, Y ⇒ Y IS X APPOSITION(X,Y)
∴ H_2 D.L. is the recipient of the N.P. in Literature in 2007.

AA$_3$ [
lex:verbaliz X⇒Y, TYPE(X)=N, TYPE=V, VERB_OF(Y,X)
∴ H_3 D. L. received the N.P. in Literature in 2007.

AA$_4$ [
reas:gen_infer X RECEIVE PRIZE ⇒ X WON PRIZE
∴ ***H*** *D. Lessing won the N. P. in Literature in 2007.*

At step 1 of the methodology, the linguistic phenomena (i.e. apposition, synonymy, verbalization and argument realization) are considered relevant to the entailment between T and H, meaning that evidence related to such aspects should be filled in to correctly judge the pair. Applying step by

[6]The symbol [...] is used as a place-holder of the non relevant parts of the sentence that we omit for brevity.

step the procedure to the phenomenon we define as argument realization, at step 2a the following general rule is added as additional premise, to provide evidence related to the phenomenon under consideration:

Entailment rule:	**temporal_argument**
Pattern:	*X Y* $\Leftrightarrow$ *Y in X*
Constraint:	*TYPE(X)= TEMPORAL_EXPRESSION(Y,X)*
Probability:	1

Then, such general rule is instantiated (*2007 Nobel Prize in Literature* $\Leftrightarrow$ *Nobel Prize in Literature in 2007*), and at step 2b the substitution in T is carried out (*Doris Lessing, recipient of the Nobel Prize (in Literature) in 2007 [...]*) to obtain an intermediate conclusion. This step represents the first inferential step of the chain that should be carried out in the reasoning process. The atomic pair $T - H_1$ is therefore composed (step 2c) and marked as *argument realization* (macro-category *syntactic*). Finally, at step 3, this pair is judged as *entailment*. Step 2 (a, b, c) is then repeated for all the phenomena individuated in that pair at step 1, till the final conclusion is derived.

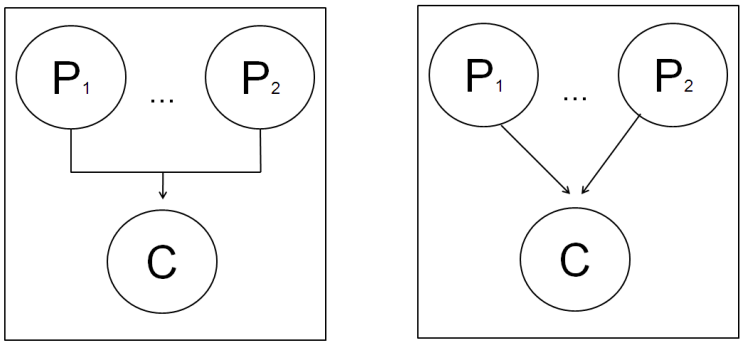
It can be the case that several phenomena are collapsed on the same tokens. For instance, in the example reported above, a chain of three phenomena should be solved to match "recipient of" with "won". In such cases, in order to create an atomic H for each phenomenon, the methodology is applied once to the first phenomenon of the chain (therefore creating the pair $T - H_i$), then it is applied again on H_i (that becomes T') to solve the second phenomenon of the chain (creating the pair $T' - H_j$); more specifically, in the example above the methodology is first applied on T for the apposition ($T - H_2$), and then, it is recursively applied on H_2 (that becomes T') to solve the verbalization ($T - H_3$). Finally, we apply it once

more on H_3 (that becomes T") to solve the general inference $(T' - H_4)$.

We experimented with the proposed methodology over a sample of pairs taken from RTE data set, and investigated critical issues arising when entailment, contradiction and unknown pairs are considered. The result is a resource, described in more details in Chapter 6, that can be profitably used to advance the comprehension of the linguistic phenomena relevant to entailment judgements.

4.3 Dependencies among atomic arguments

In Chapter 2 we explained that if an argument contains several steps of reasoning supporting all the same (final or intermediate) conclusion, the argument is said to be *convergent*. Instead, if each of the premises requires the completion by the others to derive the conclusion, the argument is said to be *non convergent*, as shown in Figure 4.1.



(a) Non convergent (b) Convergent

Figure 4.1: Arguments inferential structures

In a parallel way, in TE pairs decomposed in a set of simple premises providing the pieces of evidence needed for computational purposes, some inferential steps can independently support the final conclusion as in convergent arguments. On the contrary, some other steps of reasoning can

require information provided by other premises to infer the conclusion, as in non convergent arguments. In particular, since in our model we are decomposing T focusing on the phenomena that should be tackled to correctly infer H, we would have a convergent inferential structure in a pair when all the phenomena can independently be solved once adding the missing pieces of evidence. On the contrary, we would have a non convergent inferential structure when more than one phenomenon is instantiated on the same tokens so that the evidences concerning all these phenomena should complete each other to derive the conclusion. For instance, the inferential structure of Example 4.3 can be represented as Figure 4.2, meaning that once we have pieces of evidence supporting the correctness of the inference step related to the phenomenon we call syntactic realization, we have solved the entailment task related to that phenomenon. On the contrary, since the other phenomena relevant in the pair (i.e. apposition, verbalization, and general inference) are strongly dependent one on the other and are instantiated on the same text snippet (i.e "recipient of" - "won"), we need the completion of the missing pieces of evidence related to these phenomena to solve this sub-task of entailment.

As introduced before, the intuition underlying our proposal of decomposing the complexity of the TE task to separately tackle the phenomena relevant to inference in a pair, is motivated by the notion of meaning compositionality. According to such principle (Frege 1992 [37]), the meaning of a complex expression e in a language L is determined by the structure of e in L and by the meaning of the constituents of e in L. In a parallel way, we assume that it is possible to recognize the entailment relation of a T-H pair (i.e. to correctly judge the *entailment/contradiction* relation) only if all the phenomena contributing to such a relation are resolved. In other words, we assume that in order to validate the original argument as a whole, we need to validate all the related atomic arguments. When we

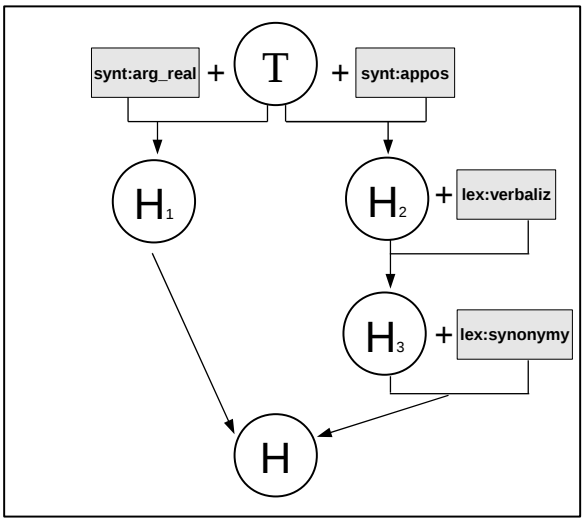
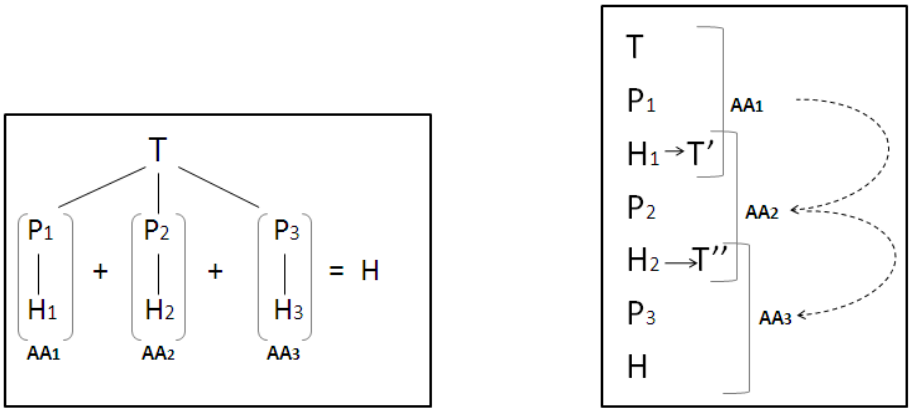


Figure 4.2: Inferential structure of Example 4.3

say "validate an argument", we mean to evaluate its correctness according to the argument evaluation criteria described in Chapter 2. To reach this goal, at each inferential step of the decomposition process the validity of the atomic argument has to be checked, and an entailment judgement has to be assigned as the output of this operation.

Once all the atomic arguments relevant to entailment in a pair have been separately solved, suitable compositional mechanisms should then be applied to combine the partial outputs to obtain a global judgement for that pair. Often, as Figure 4.2 shows, the phenomena that should be solved in a pair to correctly derive H are not independent, but interact in a complex way. Compositional mechanisms should therefore take into consideration the interactions and the dependencies of the phenomena that convey the pair meaning. For instance, if the inferential structure of the atomic arguments in a pair is convergent, sequential models of composition of partial outputs can be applied (Figure 4.3a). If it is not convergent, cascade models should be preferred (Figure 4.3b).

In the next Section, a computational framework to deal with the inferential



(a) Sequential model (b) Cascade model

Figure 4.3: Compositional models of atomic arguments

structure in TE pairs is proposed.

4.4 A component-based architecture for TE

Adopting the terminology and the definitions provided by classical logic, in the previous sections we discussed about the inferential structure of TE pairs. To take stock of the situation, let's summarize the main issues we rose and the lessons learnt:

- we proposed a model for complex arguments decomposition, to highlight the relations between the premise (i.e. T) and the conclusion (i.e. H). Implicit premises expressing both the linguistic pieces of evidence and the world knowledge required to carry out the inference task are made explicit, and added to the argument as additional premises. As a result, several atomic arguments are generated to decompose the reasoning process into a chain of inferential steps, with the goal of simplifying it;

- a categorization of the pieces of evidence required to derive H from T in TE pairs has been carried out, basing on linguistic features. The

phenomena relevant to entailment we identified define the type of linguistic evidence needed to perform an inferential step on a specific atomic argument. By translation, such phenomena classify the atomic argument itself;

- integrating Fregean meaning compositionality principle in the TE framework, we assumed a functional relation between validating the atomic arguments related to a certain argument, and validating the complex original argument as a whole. For this reason, at each inferential step of the reasoning chain, each atomic argument is checked for validity and an entailment judgement is assigned. After validating all the relevant atomic arguments in a TE pair, suitable compositional mechanisms should be applied to join the partial outputs to obtain a global judgement for that pair;

- observations on the dependencies among atomic arguments (and therefore among the phenomena relevant to derive H from T) have been pointed out, and different compositional models have been discussed.

To take full advantage of this theoretical model for computational purposes, we hypothesize a modular framework for TE, where precision- oriented components are specialized to separately carry out the inferential step related to each atomic argument. More concretely, we propose a component-based TE architecture, as a set of clearly identifiable TE modules that can be singly used on specific entailment sub-problems, and can then be combined to produce a global entailment judgement for a pair. Given a T-H pair, each component must be able to identify the phenomenon (or class of phenomena) it is build to address, and to derive an intermediate conclusion basing on the piece of evidence provided by the application of the appropriate entailment rule (atomic argument). Moreover, each component has to provide an entailment judgement for that atomic argument,

depending on its validity. Comparing the argument evaluative criteria discussed in Chapter 2 with the three-way judgements expected by TE task on T-H pairs (Chapter 3), the following correspondences come to light:

- *entailment* judgement: all the evaluation criteria are satisfied, meaning that the pair expresses a valid deductive argument, or an inductive argument with a high inductive probability;
- *contradiction* judgement: the argument is not valid, since the conclusion contradicts the premise (Criterion 2 - validity and inductive probability - is not satisfied);
- *unknown* judgement: either the inductive probability of the argument is too low to be considered a good argument (Criterion 2 is not satisfied), or the premises are not pertinent to derive the conclusion (Criterion 3 - relevance - is not satisfied).

4.4.1 TE-components expected behaviour

As introduced before, each TE-component receives a T-H pair as input, and according to our model it is expected to *i)* identify the phenomenon i it is built to address, *ii)* generate the atomic argument AA_i applying the piece of evidence related to phenomenon i that allows to derive an intermediate conclusion, and *iii)* output an entailment judgement (JUDG_i) depending on the validity of AA_i, such that:

$$\text{JUDG}_i(T,H) = \textit{neutral} \quad \text{if } i \text{ does not affect T and H (either } i \text{ is not present in the pair or it is not relevant to inference)}$$

$$\text{JUDG}_i(T,H) = \begin{cases} \textit{entailment} & \text{if } \text{AA}_i \text{ is a valid argument} \\ \textit{contradiction} & \text{if in } \text{AA}_i \text{ the conclusion (H) contradicts the premise (T)} \\ \textit{unknown} & \text{if in } \text{AA}_i \text{ the truth of H wrt T remains unknown on the basis of } i \end{cases}$$

As an example, let's suppose a TE-component which only detects entailment due to the active-passive alternation between T and H, and suppose the following T-H pairs:

T1	John painted the wall.
H1	The wall is white.
H2	The wall was painted by John.
H3	The wall was painted by Bob.

When the TE-component COMP_{a-p} is applied to the examples, according to our definition we will obtain the following results (JUDG_{a-p} is the judgement assigned with respect to the phenomenon of *active-passive alternation*):

$\text{JUDG}_{a-p}(T1, H1) = \text{unknown}$
because there is no active-passive alternation in the pair;

$\text{JUDG}_{a-p}(T1, H2) = \text{entailment}$
because the application of an active-passive rule allows to generate the conclusion (H2), meaning that AA_{a-p} is a valid argument (the entailment between T1 and H2 is preserved);

$\text{JUDG}_{a-p}(T1, H3) = \text{contradiction}$
because, although an active-passive alternation is present in the pair, the corresponding entailment rule cannot be applied, meaning that AA_{a-p} is not a valid argument (H3 contradicts T1).

More generally, we distinguish four cases in the behaviour of a TE-component COMP_i:

The neutral case, when the phenomenon i does not occur in a certain pair. We say that the TE engine COMP_i is "neutral" with respect to i, when it cannot produce any evidence either for the entailment or the contradiction between T and H.

The positive case, when the phenomenon i occurs, and the atomic argument generated through the application of the entailment rule expressing the piece of evidence needed to derive a conclusion related to i is a valid argument (i.e. AA_i contributes to establish an entailment relation between T and H). We consider *equality*, i.e. when T and H are made of the same sequence of tokens, as a special case of the positive situation.

The negative case, when the phenomenon i occurs and the atomic argument generated through the application of the entailment rule expressing the piece of evidence needed to derive a conclusion related to i is not a valid argument (T contradicts H). More specifically, negative cases may correspond to two situations: *i)* explicit knowledge about contradiction (e.g. antonyms, negation) or *ii)* a mismatch situation, where it is not possible to apply an entailment rule, and as a consequence, a certain degree of contradiction emerges from the T-H pair (see the T1-H3 pair on active-passive alternation).

The unknown case, when the phenomenon i occurs but is it not possible to prove the truth of H wrt T in AA_i, as for hyponymy/hyperonymy (*e.g. T: John is a football player*; *H2: John is a goalkeeper*).

In our model, the last three cases are defined in the same way as the judgements allowed in the TE task, while the *neutral case* is a specific possible behaviour of the component-based framework. As introduced before,

a TE-component should first recognize the phenomenon i it is built to cope with, and only if i is detected in the pair, the component will output one of the three possible judgements. It must be anticipated here that components' absence of judgement (i.e. neutral case for all the components of a set) has to be interpreted as the absence of common phenomena between T and H, resulting in the assignment of the *unknown* judgement for that pair. Even if the *neutral* and the *unknown* case could result in the assignment of the same entailment relation, from our viewpoint the components' behaviour is qualitatively different.

Summing up, in a component-based architecture, each component is in turn a TE system, that performs the TE task focusing only on a certain sub-aspect of entailment. Such components must be *disjoint* one from the other, meaning that the same atomic argument (e.g. temporal, spatial inferences) cannot be covered by more than one module: this is because in the combination phase we do not want the same phenomenon to be counted more than once.

No specific constraints are defined with respect to how such components should be implemented, i.e. they can be either a set of classifiers or rule-based modules. In addition, linguistic processing and annotation of the input data (e.g. parsing, NER, semantic role labelling) can be required by a component according to the phenomenon it considers. An algorithm is then applied to judge the entailment relation between T and H with respect to that specific aspect. Unlike similarity algorithms (e.g. word overlap, cosine similarity), with whom algorithms performing entailment are often associated in the literature, the latter are characterized by the fact that the relation on which they are asked to judge is directional.

4.4.2 Transformation-based framework

As introduced before, the application of entailment rules in atomic arguments produces a minimal transformation of the premise into an intermediate conclusion. To better approximate the argument inferential structure, we assume a transformation-based model, meaning that in order to assign the correct entailment relation to a given pair, the text T is transformed into H by means of a set of edit operations. Each inferential step of the reasoning chain is the result of the transformation of a premise into an intermediate (or final) conclusion, through the application of edit operations (i.e. insertion, deletion, substitution). Atomic edits allowed for a specific phenomenon are expressed in terms of application of entailment rules (as defined in Section 4.2.3). More specifically, in our component-based architecture, each TE-component[7] first identifies the phenomenon it is built to address, and then generates a conclusion resulting from the application of *atomic edits* to the portions of T and H expressing that phenomenon, as shown in Figure 4.4. Each single transformation (i.e. *atomic edit*) can have a different granularity, according to the category of the phenomenon that is considered. For instance, transformations relative to lexical phenomena would probably involve single words, while syntactic transformations would most likely involve manipulation of syntactic structures. An entailment judgement is then assigned to the resulting atomic argument, depending on its validity, as explained in Section 4.4.1.

According to our framework, the nature of the TE task is not modified, since each atomic argument independently solved by the TE-components keeps on being an entailment task. Suitable composition mechanisms should then be applied to combine the output of each single component to obtain a global judgement for a pair. This issue will be the topic of the

[7]In our previous papers we used to refer to TE component as *specialized entailment engines*.

next Section.

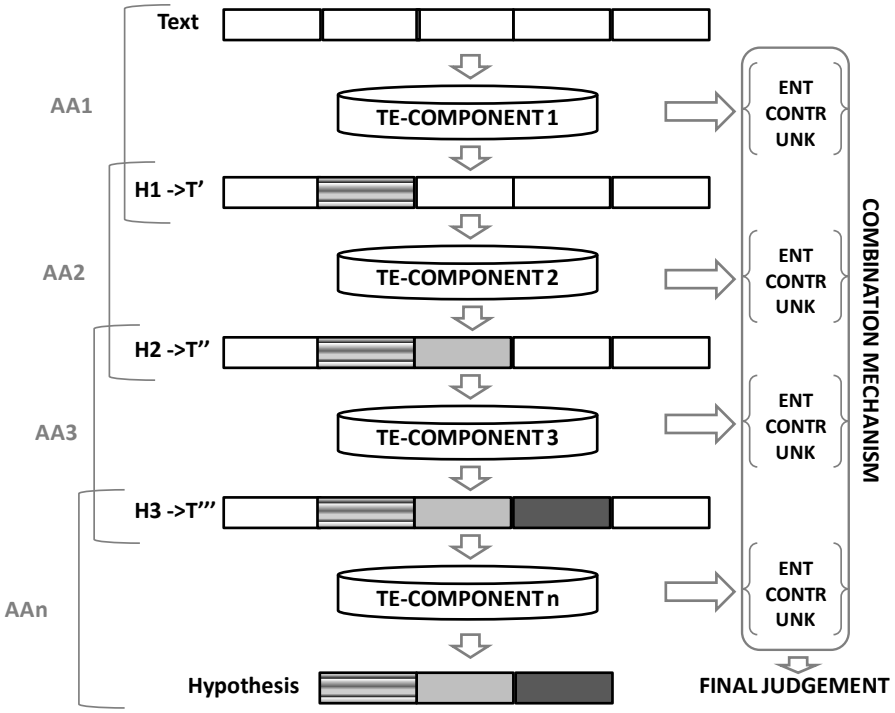


Figure 4.4: Component-based architecture

4.5 Natural Logic for TE-components definition

In the previous Section we defined the criteria that should be fulfilled in a component-based architecture, and we outlined the behaviours expected by each TE-component to be compliant with this framework. From a computational viewpoint, we need to go a step further: we need to define the combination mechanisms to join the judgements - independently provided by each component on a specific atomic argument - to obtain a global entailment judgement for a pair. To reach this goal, we take advantage of the conceptual and formal tools available from an extended model of Natural Logic (NL) (MacCartney and Manning 2009 [56]), that provides compositional operators applied on a set of well-defined semantic relations. This

model fits well in our component-based framework, and establishes clearer specifications to better formalize it.

4.5.1 Extended model of Natural Logic

Natural Logic provides a conceptual and formal framework for analysing natural inferential systems in human reasoning, without full semantic interpretation. Originating in Aristotle's syllogisms, it has been revived in the '80s in works of van Benthem (1988) [10], Sánchez Valencia (1991) [90], and Nairn *et al.* (2006) [71].

In this Section we introduce the concepts of the NL framework that we used to give shape to our component-based model, to account for natural language inference problems. In particular, in (MacCartney and Manning 2009 [56]) the authors propose a natural language inference model based on natural logic, which extends the monotonicity calculus to incorporate semantic exclusion, and partly unifies it with Nairn *et al.*'s account of implicatives. First, the authors define an inventory of *basic semantic relations* (set $\mathfrak{B}$) including representations of both containment and exclusion, by analogy with set relations[8] (shown in Table 4.1). Such relations are defined for expressions of every semantic type: sentences, common and proper nouns, transitive and intransitive verbs, adjectives, and so on. This aspect is relevant to our goals, since we would like to handle variability in natural language inference at different linguistic levels.

In $\mathfrak{B}$, the semantic containment relations ($\sqsubseteq$ and $\sqsupseteq$) of the monotonicity calculus are preserved, but are decomposed into three mutually exclusive relations: *equivalence* ($\equiv$), *(strict) forward entailment* ($\sqsubset$), and *(strict) reverse entailment* ($\sqsupset$). Two relations express semantic exclusion: *negation*($\hat{}$), or exhaustive exclusion (analogous to set complement), and *alter-*

[8] In a practical model of informal natural language inference, they assume the non-vacuity of the expressions.

symbol	name	example	set theoretic definition
$x \equiv y$	equivalence	couch $\equiv$ sofa	$x = y$
$x \sqsubset y$	forward entailment	crow $\sqsubset$ bird	$x \subset y$
$x \sqsupset y$	reverse entailment	European $\sqsupset$ French	$x \supset y$
$x \hat{} y$	negation	human $\hat{}$ nonhuman	$x \cap y = 0 \wedge x \cup y = U$
$x \mid y$	alternation	cat $\mid$ dog	$x \cap y = 0 \wedge x \cup y \neq U$
$x \smile y$	cover	animal $\equiv$ nonhuman	$x \cap y \neq 0 \wedge x \cup y = U$
$x \# y$	independence	hungry # hyppo	(all other cases)

Table 4.1: Set $\mathfrak{B}$ of basic semantic relations (MacCartney and Manning 2009 [56])

nation ($|$) or non-exhaustive exclusion. Another relation is *cover* ($\smile$), or non-exclusive exhaustion; finally the *independence relation* (#) covers all other cases (non-equivalence, non-containment, non-exclusion, and non-exhaustion). The relations in $\mathfrak{B}$ are mutually exclusive, and it is possible to define a function $\beta(x, y)$ that maps every ordered pairs of non vacuous expressions to the unique relation in $\mathfrak{B}$ to which it belongs.

Furthermore, a model to join ($\bowtie$) semantic relations is provided, as shown in Table 4.2. It could happen that the result of joining two relations is not a relation in $\mathfrak{B}$, but the union of such relations (specifically $\bigcup\{\equiv, \sqsubset, \sqsupset, |, \#\}$), meaning that the relation is not determined (refer to MacCartney and Manning 2009 [56] for further details, and for explanations on the theoretical foundation of the model). The total relation, notated as $\bullet$, is the relation that contains all pairs of (non-vacuous) expressions and conveys zero information about them.

After providing the basic definitions of the building blocks of their model of natural language inference, MacCartney and Manning (2009) [56] describe a general method for establishing the semantic relations between a premise p and an hypothesis h. The steps are as follows:

1. Find a sequence of atomic edits (i.e. deletion, insertion, or substitution of a subexpression) $< e_1, ..., e_n >$ which transforms p into h

$\bowtie$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$\equiv$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$\sqsubset$	$\sqsubset$	$\sqsubset$	$\equiv\sqsubset\sqsupset\mid\#$	$\mid$	$\mid$	$\sqsubset\wedge\mid\smile\#$	$\sqsubset\mid\#$
$\sqsupset$	$\sqsupset$	$\equiv\sqsubset\sqsupset\mid\#$	$\sqsupset$	$\smile$	$\sqsupset\wedge\mid\smile\#$	$\mid$	$\sqsupset\smile\#$
$\wedge$	$\wedge$	$\smile$	$\mid$	$\equiv$	$\sqsupset$	$\sqsubset$	$\#$
$\mid$	$\mid$	$\sqsubset\wedge\mid\smile\#$	$\mid$	$\sqsubset$	$\equiv\sqsubset\sqsupset\mid\#$	$\sqsubset$	$\sqsubset\mid\#$
$\smile$	$\smile$	$\smile$	$\sqsupset\wedge\mid\smile\#$	$\sqsupset$	$\sqsupset$	$\equiv\sqsubset\sqsupset\mid\#$	$\sqsupset\smile\#$
$\#$	$\#$	$\sqsubset\smile\#$	$\sqsupset\mid\#$	$\#$	$\sqsupset\mid\#$	$\sqsubset\smile\#$	$\bullet$

Table 4.2: Join table for relations in $\mathfrak{B}$ (MacCartney and Manning 2009 [56])

2. For each atomic edit e_i :

   (a) Determine the lexical semantic relation $\beta(e_i)$;

   (b) Since $\beta(e_i)$ depends on properties of the context of the expression in which e is applied, compute the projection of $\beta(e_i)$ upward through the semantic composition tree of the expression, while respecting the monotonicity properties of each node along the path;[9]

3. Join atomic semantic relations across the sequence of edits.

This model has been implemented in software as the NatLog system, and has been evaluated on both *i)* on the FraCaS test suite (Cooper *et al.* 1996 [35]), and *ii)* on the RTE-3 test suite (Giampiccolo *et al.* 2007 [40]). NatLog obtained better results (MacCartney and Manning 2007 [54], MacCartney and Manning 2008 [55]) on the first test suite with respect to RTE data, since the latter contains a variety of types of inference (e.g. paraphrase, temporal reasoning, relation extraction) that NatLog is not designed to address. In (Chambers *et al.* 2007 [20]) strategies of hybridizing the model with broad-coverage RTE systems have been experimented.

[9]More details on how this is performed are provided in (MacCartney and Manning 2009) [56].

In our framework we take advantage of this model, adopting both the set of semantic relations and the mechanisms for their combination. A step further, we provide an operational definition of atomic edits, in terms of application of entailment rules expressing the knowledge of a certain linguistic phenomenon (Section 4.5.3).

4.5.2 Defining TE-components using NL relations

As introduced in Section 4.3, the proposed framework assumes Fregean meaning compositionality, meaning that we hypothesize that the correct entailment judgement (JUDG) can be assigned to a $T - H$ pair combining the entailment relations (equivalent to the semantic relations described in Section 4.5.1) separately assigned to the different atomic arguments generated to derive H from T. In other words, given $\text{JUDG}_i(T - H)$, the relation assigned to the atomic argument AA_i, we assume that:

$$\text{JUDG}(T, H) = COMB_{i=1}^{n}[\text{JUDG}_i(T, H)] \tag{4.6}$$

where i potentially ranges over all the phenomena involved in textual entailment, and COMB is the composition function. According to our initial assumptions, and in line with the NL approach described in Section 4.5.1, we expect the possibility to assign to each atomic argument derived from a T-H pair one of these relations. In the transformation-based framework we assume (described in Section 4.4.2), the assignment of such relations is the result of the application of edit operations to the portions of T and H expressing the phenomenon under consideration. The correct combination of all the relations provided for the atomic arguments in a pair would then result in the assignment of the final entailment judgement to the pair.
Compliant with the definitions provided in Section 4.4.1, Natural Logic allows us to refine the possible behaviours of a TE-component in terms of more fine-grained judgements, i.e. the set of basic semantic relations:

- the *neutral case*, i.e. when the phenomenon i does not occur in a certain pair. With respect to phenomenon i, a relation of *independence* exists between T and H ($T\#H$);
- the *positive case*, i.e. when the atomic argument AA_i is a valid argument (it contributes to establish an entailment relation between T and H). Both the relations of *equivalence* ($T \equiv H$) and *forward entailment* ($T \sqsubset H$) fall within this case;
- the *negative case*, i.e. when the atomic argument AA_i is not a valid argument (it contributes to establish a contradiction relation between T and H). Both the relations of *negation* ($T \hat{} H$) and *alternation* ($T \mid H$) fall within this case;
- the *unknown case*, when it is not possible to prove the truth of H wrt T on the basis of i. Both the relations of *cover* ($T \smile H$) and *reverse entailment* ($T \sqsupset H$) fall within this case.

4.5.3 Entailment Rules and Atomic Edits

In our transformation-based framework, atomic edits are applied to sub portions of T and H expressing a certain linguistic phenomenon, and their granularity is defined by the linguistic phenomenon they describe. More specifically, we define the allowed transformations (i.e. atomic edits) for a certain linguistic phenomenon through a set of entailment rules for that specific phenomenon, as explained in Section 4.2.3.

Supposing to have a repository of all the entailment rules expressing the knowledge about the linguistic phenomena relevant to inference, we could associate an entailment relation both to the correct and to the incorrect application of the rule. For instance, the correct instantiation of the entailment rule for *active/passive alternation* expressed as:

Entailment rule:	**active/passive alternation**
Pattern:	*X V1 Y* $\Leftrightarrow$ *Y V2 by X*
Constraint:	*SAME_LEMMA(V1,V2)*
	TYPE(V1)= ACTIVE_FORM
	TYPE(V2)=PASSIVE_FORM
Probability:	1

and instantiated as e.g. T: *John painted the wall* $\Leftrightarrow$ H1: *The wall was painted by John* maintains the equivalence relation between T and H (T $\equiv$ H), and the pair (T, H1) should be marked as *entailment*. The wrong instantiation of the same rule as in H2: *The wall was painted by Bob* produces an alternation relation (T | H), and the *contradiction* judgement should be assigned to the pair (T, H2). Following the same criteria, for *hyponymy/hyperonymy* the entailment rule is expressed as:

Entailment rule:	**hyponymy**
Pattern:	$X \Rightarrow Y$
Constraint:	*HYPONYMY(X,Y)*
Probability:	1

and instantiated as e.g. T: *John is a football player* $\Rightarrow$ H1: *John is an athlete*. According to this phenomenon, a forward entailment relation exists between T and H ($T \sqsubset H$) and the pair (T, H1) should be marked as *entailment*. Instead, the inversion of the directional entailment rule as in H2: *John is a goalkeeper* produces a reverse entailment relation between T and H ($T \sqsupset H$), and the pair (T, H2) should be marked as *unknown*.

4.6 TE-components combination

In Sections 4.5.2 and 4.5.3 we have described in details the elements of our framework, defining the TE-components and their possible behaviours in terms of entailment relations to be assigned to the linguistic phenomena relevant to inference in a given pair. The inference task is therefore decomposed into a sequence of atomic inference problems, separately solved by a set of disjoint precision-oriented modules, each of which outputs *i)* the entailment relation corresponding to the processed linguistic phenomenon in a pair, and *ii)* the set of transformations between T and H allowed by the application of entailment rules for that specific phenomenon. In this Section we go a step further, taking advantage of the mechanisms of relation composition provided by the extended model of NL presented in Section 4.5.1, to combine the outputs of the TE-components to obtain a global judgement for a pair.

4.6.1 Combination based on Natural Logic

Table 4.2 (Section 4.5.1) describes the relations resulting from joining the atomic semantic relations across the sequences of edits, according to the model presented in (MacCartney and Manning 2009 [56]). Adopting this strategy in our component-based framework, we compose step by step the entailment relations separately assigned by each component to determine the global entailment relation for a pair.

Relation composition is deterministic, and in general it follows intuitive rules (e.g. $\equiv$ composed with $\equiv$ yields $\equiv$, $\sqsubset$ composed with $\sqsubset$ yields $\sqsubset$). At each step, the result may be either a basic entailment relation, or the union of such relations, with larger unions conveying less information about entailment (i.e. every union relation which results from joining relations in $\mathfrak{B}$ contains #, and thus can be approximated by #). As a drawback,

it must be noticed that composition tends to degenerate towards # both because composing # with any relation yields that relation, and because composing a chain of randomly-selected relations tends towards # as the chain grows longer. In our framework, such relation is assigned if the TE component is *neutral* with respect to a certain pair, meaning that the phenomenon is it built to deal with is not present. In this case, such relation is not counted in the composition phase[10].

4.6.2 Order of composition

The fact that the TE-components are disjoint does not guarantee that they are independent, which means that the order of their application does affect the final result. For instance, considering the pair T: *John painted the wall* - H: *The wall was coloured by John*, it seems difficult to apply the active-passive transformation before the lexical transformation between "paint" and "colour" has been carried out. We therefore assume a *cascade* of disjoint TE-components, where each component takes as input the output of the previous one, defined as the set of edit transformations from T to H_i. The order in which the TE-components are run does not correspond to sentence order, but is defined through linguistically-motivated heuristics; this ordering defines a path from T to H through intermediate forms. As a first approximation, we first run the engines whose transformations apply to single tokens, such as lexical phenomena (e.g. synonymy, hypernymy), then the engines involving structures, like syntactic phenomena (e.g. active/passive alternation, argument realization) and discourse phenomena (e.g. zero anaphora), and finally reasoning (e.g. spatial, temporal reasoning).

With respect to the final entailment judgement, if the combination of the

[10]If all the components output #, it means that no phenomena are in common between T and H, i.e. the relation is *unknown*.

relations separately assigned to the different linguistic phenomena present in T and H is either $\equiv$ or $\sqsubset$, the *entailment* judgement is assigned to the T-H pair. On the contrary, if it is either $\hat{}$ or $|$ the *contradiction* judgement is assigned, while if it is either $\sqsupset$, $\smile$, or # the *unknown* judgement is assigned.

4.6.3 Experimenting NL combination mechanisms on RTE pairs

In the TE component-based framework we propose, we suppose to have a set of TE-components covering the most frequent phenomena relevant to inference, and behaving as defined in Section 4.5.2. As an exercise, we run them in the order hypothesized before on an entailment pair, on a contradiction pair, and on an unknown pair extracted from RTE-5 test set (respectively, pairs 123, 408 and 422) (Bentivogli *et al.* 2009 [14]). Tables 4.3, 4.4 and 4.5 show the procedure for combining the semantic relations obtained by the TE-components on the example pairs, basing on NL combination mechanisms (MacCartney and Manning 2009 [56]). In each table, only the output of the components built to deal with the phenomena relevant to inference in that specific pair (i.e. non-neutral components) is presented. All the other components of the set are expected to be neutral (expected output = #), and their judgement is not taken into account in the combination phase.

On the entailment pair presented in Table 4.3, four components should be activated, namely those dealing with coreference, nominalization, modifiers, and paraphrase. Each of them is expected to carry out atomic edits (i.e. insertion, deletion or substitution) on the portions of T and H expressing the phenomena detected, applying the corresponding entailment rules. As output, each component provides both the entailment relation assigned to that operation (JUDG_i), and an intermediate form of H (intermediate conclusion) expressing the instantiation of the rule in that specific pair (H_i). The entailment relation produced by each engine is then combined

	Text snippet (pair 123)		Atomic edit/rule	Component/ Phenomena	$judg_i$	$judg_{COMB}$
T	[...] **Susan Boyle**, 47, wowed judges alike when she **performed on the television contest "Britain's got Talent."** [...]					
	H_1	**Susan Boyle** performed on the television contest "Britain's Got Talent." [...]	**x⇔y coref(x,y)**	**disc:coref**	≡	≡
	H_2	Susan Boyle **is a performer** of the television contest "Britain's Got Talent."	**x⇒ y verbal_of(y,x**	**lexsynt: verb_nom**	≡	≡
	H_3	Susan Boyle is a performer of **the contest** "Britain's Got Talent."	**x y⇒y modif(x,y)**	**synt:modif**	⊏	⊏
	H_4	Susan Boyle is **a contestant** on "Britain's Got Talent."	*a performer on a contest ⇒ a contestant*	**lexsynt: paraphrase**	≡	⊏
H	Susan Boyle is a contestant on "Britain's Got Talent."					⊏

Table 4.3: Application of the NL composition methodology to an *entailment* pair.

with the one assigned by the previous component in the chain, following the semantic relation combination scheme described in Table 4.2 (JUDG_{COMB}). Finally, the last combination step produces the judgement to be assigned to the pair. For instance, in the first example (Table 4.3) the final relation is ⊏, therefore the pair is judged as *entailment*.

On the contradiction pair presented in Table 4.4, three components should be activated, namely those dealing with semantic opposition, argument realization, and apposition. Following the same procedure described for the previous example, the last combination step produces | as final relation, meaning that the pair should be judged as *contradiction*.

On the last example we describe, i.e. the unknown pair presented in

		Text snippet (pair 408)	Atomic edit/rule	Component/ Phenomena	$judg_i$	judg $_{COMB}$
T		**Mexico's new president, Felipe Calderon,** seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]				
	H_1	Mexico's **outgoing** president, Felipe Calderon [...]	**x ⇎ y sem_opp(x,y)**	**lex:sem_opp**	|	|
	H_2	The outgoing president **of**, **Mexico** Felipe Calderon [...]	**x's y ⇒y of x**	**synt:arg_realiz**	≡	|
	H_3	**Felipe Calderon is** the outgoing **President** of Mexico.	**x,y⇒y is x apposit(y,x)**	**synt:apposit**	≡	|
H		Felipe Calderon is the outgoing President of Mexico.				|

Table 4.4: Application of the NL composition methodology to a *contradiction* pair.

Table 4.5, four components should be activated, namely those dealing with the phenomenon we call coordination, general reasoning, modifier and hyponymy. Again, we apply the procedure described for the previous examples, and in this case the last combination step produces a union of relations that tends towards #, meaning that the pair should be judged as *unknown*. It must be noticed, however, that in this example the order of application of the components does not follow the one we hypothesized in Section 4.6.2, since the step related to the general inference *Gillette, known for brands as x ⇒ Gillette manufactures x* must precede the others to proceed in the inferential chain. Experimenting this methodology for the combination of semantic relation on a sample of RTE pairs, for some examples we came up against the problem pointed out in (MacCartney and Manning 2009 [56]), i.e. the fact that composing a chain of relations tends towards # as the chain grows longer, conveying no information about the entailment.

	Text snippet (pair 422)		Atomic edit/rule	Component/ Phenomena	$judg_i$	judg $_{COMB}$
T	[...] **Gillette, known for brands such as** Gillette razors, **Oral B dental care**, and Duracel batteries, has had growing problems [...]					
	H_1	Gillette, known for brands such as Oral B dental care.	**x, y, z⇒x**	**synt:coord**	≡	≡
	H_2	Gillette **manufactures** Oral B dental care.	**x**, *known for brands such as* **y** ⇒ **x** *manufactures* **y**	**reas: gen_infer**	≡	≡
	H_3	Gillette manufactures **dental care** (products).	**x y⇒y modif(x,y)**	**synt:modif**	⊏	⊏
	H_4	Gillette manufactures **toothpaste**.	**x =?⇒y hypon(x,y)**	**lex:hypon**	⊐	≡⊏⊐ |#
H	Gillette manufactures toothpaste.					∼ #

Table 4.5: Application of the NL composition methodology to a *unknown* pair.

4.7 Conclusion

Progressively abandoning the parallelism with logical arguments that up to now we used to motivate and position our proposal from a theoretical viewpoint, in this Chapter we started to direct our attention towards more computational aspects of the framework. In particular, we focused on the definition and formalization of an architecture for component-based Textual Entailment, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. We took advantage of the conceptual and formal tools available from an extended model of Natural Logic (NL) to define clear strategies for their combination, in a transformation-based framework. With respect to the model described in (MacCartney and Manning 2009 [56]) in which a lot of effort

is made to establish the proper projectivity signatures for a broad range of quantifiers, implicative and factives, and other semantic relations, our work is less fine-grained, since it relies on the expressivity of the entailment rules to model a certain linguistic phenomenon. On the other hand, as far as a linguistic phenomenon can be expressed through entailment rules it can be modelled in our framework, guaranteeing a broader coverage on RTE problems.

In the next Chapter, we experiment the feasibility of the component-based TE framework we proposed, adopting a modular architecture that accounts for the properties of the components described above.

Chapter 5

Implementation of TE-components based on EDITS architecture

To experiment the feasibility of the component-based TE framework proposed in this Thesis, we take advantage of the flexible and modular architecture of the EDITS system (Kouylekov and Negri 2010 [49]) for the implementation of a set of TE-components. In this Chapter we describe how these modules have been designed, and the preliminary experiments we carried out to evaluate them on RTE data.

5.1 Introduction

In Chapter 4 we defined an architecture for component-based Textual Entailment, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. To better approximate the argument inferential structure, we assumed a transformation-based model, meaning that to assign the correct entailment relation to a given pair, the text T is transformed into H by means of a set of edit operations. Summing up, in our component-based architecture each TE-component first identifies the phenomenon it is built to address, and then generates a conclusion resulting from the application of atomic edits to

the portions of T and H expressing that phenomenon. Each single transformation (i.e. atomic edit) is allowed by the application of entailment rules for that specific phenomenon, that can have a different granularity according to the category of the phenomenon that is considered. An entailment judgement is then assigned depending on the validity of the resulting atomic argument. According to our framework, the nature of the TE task is not modified, since each atomic argument independently solved by the TE-components keeps on being an entailment task.

To experiment the feasibility of the component-based TE architecture, we take advantage of the flexible and modular architecture of the EDITS system (Edit Distance Textual Entailment Suite), an open-source software package for recognizing TE[1] developed by the HLT group at FBK (Kouylekov and Negri 2010 [49], Negri *et al.* 2009 [72]). EDITS provides a basic framework for a distance-based approach to the task, with a highly configurable and customizable environment to experiment with different algorithms (Section 5.2). Taking advantage of its potential in terms of extensions and integrations with new algorithms and resources, we used EDITS as the basic architecture for the implementation of a set of TE-components (Section 5.3). The design of each component (e.g. the linguistic preprocessing required on the input pairs, the knowledge resources, and the algorithm) strongly depends on the specific phenomenon it should detect and express an entailment judgement about. In line with the architecture definition we provided in the previous Chapter, the same inference type is not covered by more than one component. To assess the capabilities of the TE-components we designed, we carried out some experiments on RTE data sets (Section 5.4).

After independently testing each module, suitable composition mechanisms should then be applied to combine the output of each single com-

[1] `http://edits.fbk.eu/`

ponent to obtain a global judgement for a pair. In Section 5.5, simple combination strategies are experimented, namely weighted linear composition and sequential composition of the partial judgements.

More generally, a preliminary evaluation of this framework has been carried out in our participations to RTE campaigns (in particular in RTE-4, Cabrio *et al.* 2008 [18]), on standard RTE data sets provided by the organizers of the challenges. In this context, it is also worth mentioning the work of Wang and Neumann (2008) [94] (see Chapter 3), that provides an empirical evidence of the benefit of developing a modular approach to recognize TE. In particular, their system is composed of three specialized RTE-modules: *i)* to tackle temporal expressions; *ii)* to deal with other types of NEs; *iii)* to deal with cases with two arguments for each event. Besides these precision-oriented modules, two robust but less accurate backup strategies are considered, to deal with not yet covered cases. In the final stage, the results of all specialized and backup modules are joint together, applying a weighted voting mechanism.

5.2 The EDITS system

As introduced before, to experiment the feasibility of the component-based TE architecture, we take advantage of the flexible and modular architecture of the EDITS system (Edit Distance Textual Entailment Suite) (Kouylekov and Negri 2010 [49], Negri *et al.* 2009 [72]). EDITS is a TE system based on edit distance algorithms, and computes the distance between T and H as the cost of the edit operations (i.e. insertion, deletion and substitution) that are necessary to transform T into H. EDITS requires that the following modules are defined in a configuration file:

- an edit distance algorithm: e.g. Token Edit Distance - a token-based version of the Levenshtein distance algorithm, with edit operations

defined over sequences of tokens of T and H - and Tree Edit Distance - an implementation of the algorithm described in (Zhang and Shasha 1990 [102]), with edit operations defined over single nodes of a syntactic representation of T and H;

- a cost scheme for the edit operations: it explicitly associates a cost (a positive real number) to each edit operation applied to elements of T and H (it is defined as XML files). According to the algorithm used, operations are carried out either on words (with Token Edit Distance) or over nodes in a dependency tree representation (with Tree Edit Distance). In the creation of new cost schemes, users can express edit operation costs, and conditions over the words/nodes using a meta-language based on a lisp-like syntax;

- a cost optimizer (optional): to adapt cost schemes to the specific data set. The optimizer is based on cost adaptation through genetic algorithms, as proposed in (Mehdad 2009 [64]).

- a set of rules expressing either entailment or contradiction: to provide knowledge (e.g. lexical, syntactic, semantic) about the probability of entailment or contradiction between elements of T and H. Rules are invoked by cost schemes to influence the cost of substitutions between elements of T and H. Typically, the cost of the substitution between two elements A and B is inversely proportional to the probability that A entails B.

Figure 5.1 shows EDITS architecture and work flow. The input of the system is an entailment corpus represented in the EDITS Text Annotation Format (ETAF), a simple XML internal annotation format. ETAF is used to represent both the input T-H pairs, and the entailment and contradiction rules, and allows to represent texts at different levels (i.e. as sequences of

tokens with their associated morpho-syntactic properties, or as syntactic trees with structural relations among nodes). Given a configuration file and an RTE corpus annotated in ETAF, the training procedure is run to learn a model (i.e. the threshold to separate positive from negative pairs). Given a model and an un-annotated RTE corpus as input, the test procedure produces a file containing for each pair: *i)* the decision of the system (YES, NO), *ii)* the confidence of the decision, *iii)* the entailment score, *iv)* the sequence of edit operations made to calculate the entailment score.

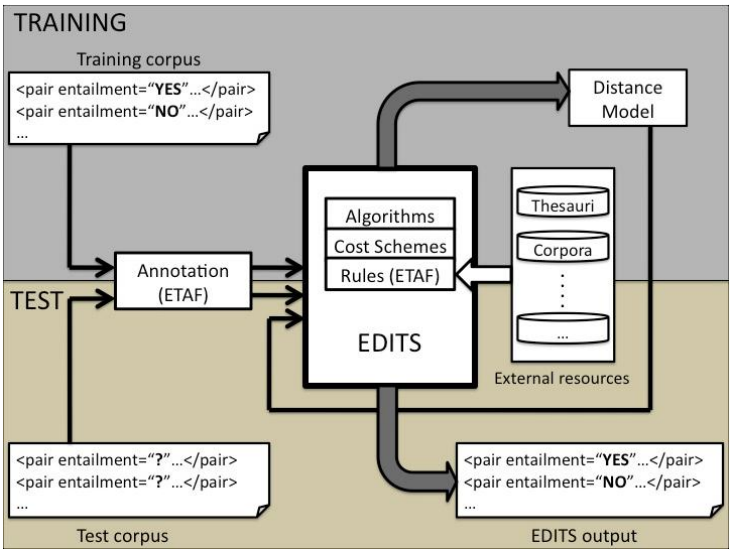


Figure 5.1: EDITS architecture and work-flow

Given the modular architecture, and the fact that each module can be easily configured by the user as well as the system parameters, we considered the EDITS system (version 1.0) as a suitable framework to experiment our component-based architecture. Moreover, EDITS can work at different levels of complexity, depending on the linguistic analysis carried on over T and H, although transformations are allowed on nodes. It also allows the integration of additional linguistic processors and semantic resources. For the implementation of our TE-components, we considered Tree Edit

Distance as the edit distance algorithm over the dependency trees of T and H obtained using Xerox Incremental Parser (XIP) (Ait-Mokhtar *et al.* 2002 [1]).

5.3 EDITS-based TE-components

With respect to the theoretical model we proposed in Chapter 4, where the output of each component is a semantic relation as defined by the extended model of Natural Logic (MacCartney 2009 [53]), for our implementation we propose a simplified version based on distance, as allowed by EDITS. We therefore assume that the edit distance ED(T-H) related to a pair can be usefully decomposed as the combination of the distances related to the different phenomena involved in the entailment relation between T and H:

$$\text{ED}(T - H) = COMB_{i=1}^{n}[\text{ED}_i(T, H)] \tag{5.1}$$

where i potentially ranges over all the phenomena involved in TE, and COMB is the composition function. Each TE-component is therefore expected to provide a distance concerning only the phenomenon i (or the category of phenomena) it is built to address. Adapting the expected behaviours of the TE-components we hypothesized in Chapter 4 to the distance-based approach, for a T-H pair each component (COMP_i) should provide a distance ED_i(T-H) as defined:

$$\text{ED}_i(T, H) = 0 \quad \text{if } i \text{ does not affect T and H (either } i \text{ is not present in the pair or it is not relevant to inference)(}\textit{neutral behaviour}\text{)}$$

$$\text{ED}_i(T, H) = \begin{cases} 0 < d \leq t_i & \text{if } \text{AA}_i \text{ is a valid argument(}\textit{positive behaviour}\text{)} \\ > t_i & \text{if in } \text{AA}_i \text{ the conclusion (H) contradicts the premise (T)} \\ & \text{(}\textit{negative behaviour}\text{)} \\ > t_i & \text{if in } \text{AA}_i \text{ the truth of H wrt T remains unknown on the} \\ & \text{basis of } i \text{ (}\textit{unknown behaviour}\text{)} \end{cases}$$

where the threshold t_i separates the entailment and the contradiction/ unknown cases due to the phenomenon i. Since EDITS is designed to recognize TE according to the two-way judgement task, the negative and the unknown behaviours are collapsed, resulting in a distance above the threshold (i.e. no entailment). The behaviour of each component is disjoint from the others, meaning that more than one module does not cover the same phenomenon in the data set.

In EDITS, the creation of a TE-component is done by modelling the basic modules described in Section 5.2 (algorithms, cost schemes, optimizer, and rules) according to the expected behaviour defined above, through an XML configuration file.

5.4 Testing TE-components on RTE data sets

This Section describes the experimental setup, both in terms of the TE-components we implemented basing on EDITS architecture (Section 5.4.1), and the results we obtained evaluating them on RTE-5 data sets (Section 5.4.2).

5.4.1 Implemented TE-components

The feasibility of the component-based approach has been experimented using three TE-components, designed according to the criteria described in Section 5.3. Such components address three different categories of phenomena relevant to inference, namely: *i)* negation and antonymy (EDITS$_{NEG}$); *ii)* coreference (EDITS$_{COREF}$); and *iii)* lexical similarity (EDITS$_{LEX}$). The decision to focus on these phenomena is motivated by various reasons: *i)* frequency of the considered phenomena (as will be showed in more details in Chapter 6), *ii)* importance of contradiction detection; and *iii)* the fact that the current version of EDITS allows to carry out edit operations on

simple nodes only, and not on subtrees (i.e. the tree edit distance algorithm works on the dependency structures of T and H, but at the moment it is not possible to insert, substitute or delete entire subtrees). The main characteristics of the implemented components are the following:

EDITS$_{NEG}$: this component sets specific costs for edit operations on negative polarity items. The underlying intuition is that assigning high costs to these operations should prevent the system from assigning positive entailment to a T-H pair in which one of the two fragments contradicts the other. A pre-processing module marks as negated the head of direct licensors of negation, such as overt negative markers (*not*, and the bound morpheme *n't*), of negative quantifiers (*no*, *nothing*), and of strong negative adverbs (*never*). Moreover, a set of contradiction rules (28,890 rules) created extracting the terms connected by the *antonym* relation in WordNet are used as source of knowledge with respect to phenomenon of antonymy.
For instance, given Example 5.2 (pair 588, RTE-5 test pair):

(5.2) T: *Sam Brownback is perplexed. The U.S. Senator from Kansas and Presidential candidate is a Republican whose politics - he is against marriage for gay people, he is against abortion, and he has a clean image in a party tainted by scandal - should speak favorably to the party's base. [...]*

H: *Sam Brownback is not a Republican.*

during the pre-processing phase, both the direct licensor of negation *not* and its syntactic head *be*, that are present in H, are annotated (basing on the dependency representation of the pair provided by XIP parser). They are represented in the ETAF format by the truth condition of the attributes "neg" for the negation and "IsNeg" for its head, as follows:

```
<node id="437-439:10">
   <word id="437-439:10">
```

```
        <attribute name="lemma">be</attribute>
        <attribute name="IsNeg">TRUE</attribute>
        <attribute name="token">is</attribute>
        <attribute name="pos">VERB</attribute>
        <attribute name="wnpos">v</attribute>
    </word>
</node>
<node id="440-443:12">
    <word id="440-443:12">
        <attribute name="lemma">not</attribute>
        <attribute name="neg">TRUE</attribute>
        <attribute name="token">not</attribute>
        <attribute name="pos">ADV</attribute>
        <attribute name="wnpos">r</attribute>
    </word>
</node>
<node id="446-456:17">
    <word id="446-456:17">
        <attribute name="lemma">republican</attribute>
        <attribute name="token">Republican</attribute>
        <attribute name="pos">NADJ</attribute>
        <attribute name="wnpos">a</attribute>
    </word>
</node>
```

After launching EDITS$_{NEG}$ on the data, the following cost-scheme is applied, setting a high cost to the substitution of a negated node with a non-negated node in the same syntactic position, and with the same part of speech (these constraints for the application of the rule are expressed as conditions):

```
<substitution name="sub-negated-lemma-hyp">
    <condition>(is-word-node A)</condition>
    <condition>(is-word-node B)</condition>
    <condition>(set posa (a.wnpos (word B)))</condition>
    <condition>(not (null posa))</condition>
    <condition>(set posb (a.wnpos (word A)))</condition>
```

```
   <condition>(not (null posb))</condition>
   <condition>(equals posa posb)</condition>
   <condition>(attribute "IsNeg" (word B))</condition>
      <cost>2.2</cost>
</substitution>
```

The following edit operation is therefore carried out on the nodes represented above, and the "sub-negated-lemma-hyp" cost-scheme is correctly applied:

```
<operation type="substitution" scheme="sub-negated-same-lemma-hyp" cost="2.2">
  <source>[node [id 84-86:27] [edge-to-parent NUCL] [word [lemma be] [token is]
          [pos VERB] [wnpos v] ]]</source>
  <target>[node [id 437-439:10] [edge-to-parent NUCL] [word [IsNeg TRUE] [lemma
          be]  [pos VERB] [token is] [wnpos v] ]]</target>
</operation>
```

EDITS$_{COREF}$: this component sets low costs for edit operations among co-referent terms, and high costs for operations between two terms that do not co-refer. During the preprocessing phase, the coreference module internal to the XIP parser identifies the Named Entities and annotates both the intra-sentential (in T and in H separately) and inter-sentential (in the pair) co-referent ones.
For instance, given Example 5.3 (pair 104, RTE-5 test pair):

(5.3) T: *Leftist Mauricio Funes of El Salvador's former Marxist rebel FMLN party has won the country's presidential election. He defeated his conservative rival, the Arena party's Rodrigo Avila, who has admitted defeat. [...]*

H: *In El Salvador Mauricio Funes has defeated Rodrigo Avila.*

during the pre-processing phase, the Named Entities and the inter-sentential and intra-sentential co-referent terms are annotated, meaning for instance

both *Mauricio Funes*, the pronoun *he* in T, and *Mauricio Funes* in H, all referring to the same person. They are represented in the ETAF format using the attribute "coref", and the id of the node where the entity appeared for the first time (used as reference id). Therefore, in T:

```
<node id="8-22:6">
    <word id="8-22:6">
       <attribute name="lemma">Mauricio Funes</attribute>
       <attribute name="token">Mauricio Funes</attribute>
       <attribute name="pos">NP</attribute>
       <attribute name="wnpos">n</attribute>
       <attribute name="coref">8-22:6</attribute>
   </word>
</node>
</node>
<node id="129-132:10">
   <word id="129-132:10">
      <attribute name="lemma">he</attribute>
      <attribute name="token">his</attribute>
      <attribute name="pos">PRON</attribute>
      <attribute name="coref">8-22:6</attribute>
   </word>
</node>
```

and then in H:

```
<node id="609-623:9">
   <word id="609-623:9">
      <attribute name="lemma">Mauricio Funes</attribute>
      <attribute name="token">Mauricio Funes</attribute>
      <attribute name="pos">NP</attribute>
      <attribute name="wnpos">n</attribute>
      <attribute name="coref">8-22:6</attribute>
   </word>
</node>
```

After launching EDITS$_{COREF}$ on the data, the following cost-scheme is applied, setting a very low cost (close to 0)[2] to the substitution of two co-

[2]We do not assign 0 to differentiate the positive from the neutral behaviour.

referent terms:

```
<substitution name="coref">
   <condition>(is-word-node A)</condition>
   <condition>(is-word-node B)</condition>
   <condition>(attribute "coref" (word A))</condition>
   <condition>(attribute "coref" (word B))</condition>
   <condition>(equals (attribute "coref" (word A)) (attribute "coref" (word B)))
   </condition>
       <cost>0.1</cost>
</substitution>
```

The following edit operation is therefore carried out on the nodes *he* and *Mauricio Funes*, that appear in the same syntactic position, and the "coref" cost-scheme is correctly applied:

```
<operation type="substitution" scheme="coref" cost="0.1">
  <source>[node [id 117-119:4] [edge-to-parent SUBJ-N] [word [lemma he] [token
          He] [pos PRON] [coref 8-22:6] ]]</source>
  <target>[node [id 609-623:9] [edge-to-parent SUBJ-N] [word [lemma Mauricio
          Funes] [token Mauricio Funes] [pos NP] [wnpos n] [coref 8-22:6] ]]
  </target>
</operation>
```

EDITS$_{LEX}$: this component addresses lexical similarity, setting low costs for substituting two terms that are highly related (i.e. that match an entailment rule). Entailment rules have been extracted from Wikipedia (namely, 58280 rules) computing the Latent Semantic Analysis (LSA) over this resource, between all possible node pairs (terms or lemmas) that appear in the RTE data set. The jLSI (java Latent Semantic Indexing) tool (Giuliano 2007 [41]) has been used to measure the relatedness between the term pairs. Pairs with low similarity have been filtered out setting a relatedness threshold (empirically estimated), keeping the ones whose second

term is entailed by the first one with a high probability. At first, we experimented this module extracting the terms connected by the *synonymy* and *hyponymy/hyperonymy* relation in WordNet, but since the coverage is quite small we decided to use Wikipedia rules instead, that have a higher coverage and contain also the previous ones. The cost of edit operations on stop-words are set to 0 and substitution of stopwords with content words is not allowed.
For instance, given Example 5.4 (pair 416, RTE-5 test pair):

(5.4) T: *Despite legislation enacted by Congress and signed into law by President Barack Obama on Wednesday, more than one-third of television stations in the United States are planning to move ahead with the transition to digital television, according to reports. [...]*

H: *TV stations are going to switch to digital.*

After launching EDITS$_{LEX}$ on the data, the following cost-scheme is applied, setting the cost of the substitution of two terms as inversely proportional to the probability of the entailment rule (extracted from Wikipedia) whose LHS matches a portion of T, and whose corresponding RHS matches a portion of H.

```
<substitution name="sub-entail1">
    <condition>(is-word-node A)</condition>
    <condition>(is-word-node B)</condition>
    <condition>(set probability (entail (word A) (word B) :wikipedia))</condition>
    <condition>(not (null probability))</condition>
       <cost>(* (- 1.1 probability) 1)</cost>
</substitution>
```

With respect to the pair reported above, the following entailment rule is applied, stating that the probability that *television* $\Rightarrow$ TV is equal to 1.

```
<rule>
  <t> <word> <attribute name="lemma">Television</attribute> </word></t>
  <h> <word><attribute name="lemma">TV</attribute></word></h>
    <probability>1.0</probability>
</rule>
```

Basing on the knowledge expressed by the rule, the following edit operation is carried out, and the "sub-entail1" cost-scheme is correctly applied:

```
<operation type="substitution" scheme="sub-entail1" cost="0.1">
  <source>[node [id 123-133:61] [edge-to-parent MOD] [word [lemma
           television] [token television] [pos NOUN] [wnpos n] ]]</source>
  <target>[node [id 513-515:4] [edge-to-parent MOD] [word [lemma TV]
           [pos NOUN] [token TV] [wnpos n] ]]</target>
</operation>
```

Since the TE-components we implemented in this experimental phase have a limited coverage with respect to the number of linguistic phenomena present in RTE data sets (as we will show in more details in Chapter 6), a backup strategy in the form of a component setting costs for the edit operations on phenomena not covered by the other components has been developed.

5.4.2 Results and error analysis

We independently run the EDITS-based TE-components described in the previous Section on RTE data sets, and we calculated the performances of the TE-components with respect to the expected behaviours described in Section 5.3. Table 5.1 reports the evaluation (Precision, Recall, Accuracy and F-measure) on RTE-5 test set with respect to the neutral behaviour. According to this task, the component should classify the pairs depending on whether it does not detect the phenomenon it is built to deal with -

True Positive (TP), since in this case, its behaviour is neutral - or if it detects it - True Negative (TN).

TE-comp.	# pairs	TP	FP	TN	FN	Prec. %	Rec. %	Acc.	F-meas.
EDITS$_{NEG}$	600	559	12	3	26	97.8	95.5	93.3	96.6
EDITS$_{COREF}$	200	117	49	29	5	70.4	95	73	80.8
EDITS$_{LEX}$	200	31	0	18	151	100	17	24.5	29.05

Table 5.1: Evaluation of the TE-components with respect to neutral behaviour.

Even if the TE-components have been run on the whole RTE-5 data-set (600 pairs training set, 600 pairs test set), we carried out the analysis for the evaluation on the whole test set only for EDITS$_{NEG}$ since the phenomena it covers are very rare, while for the other two components we analysed a sample of 200 pairs (column *tot. pairs* in Table 5.1). While EDITS$_{NEG}$ and EDITS$_{COREF}$ are good classifiers with respect to the neutral behaviour, meaning that they are able to detect the phenomenon they are built to deal with, EDITS$_{LEX}$ applies too often, even when the lexical similarity of two words in a couple is irrelevant to the inference relation in that pair.

As a second step, among the pairs in which the phenomena covered by the TE-components are relevant, we analyse how much the TE-components are able to classify if the phenomenon they detected contributes to preserve the entailment in the pair or if it is cause of contradiction/unknown. Such judgement concerns the atomic argument related to the phenomenon under consideration (the inference step related to that phenomenon only), regardless of the final judgement of the pair. To avoid to sum the mistakes deriving from the previous classification task, we give as input to the TE-components only the pairs marked as True Negative (i.e. the phenomenon is present and its presence has been detected by the system). Table 5.2 and 5.3 show the evaluation of the TE-components with respect to the positive and negative behaviours.

TE-comp.	# pairs	TP	FP	TN	FN	Prec. %	Rec. %	Acc.	F-meas.
EDITS$_{NEG}$	3	0	0	3	0	100	100	100	100
EDITS$_{COREF}$	29	11	4	2	12	73	47.8	44.8	59.3
EDITS$_{LEX}$	18	18	0	0	0	100	100	100	100

Table 5.2: Evaluation of the TE-components with respect to positive behaviour.

TE-comp.	# pairs	TP	FP	TN	FN	Prec. %	Rec. %	Acc.	F-meas.
EDITS$_{NEG}$	3	3	0	0	0	100	100	100	100
EDITS$_{COREF}$	29	2	12	11	4	14	33.3	44.8	19.7
EDITS$_{LEX}$	18	0	0	18	0	100	100	100	100

Table 5.3: Evaluation of the TE-components with respect to negative behaviour.

While the phenomenon of coreference can contribute to both entailment and contradiction judgements (we will discuss it in more details in Chapter 6), with respect to the phenomena covered by the other two components Tables 5.2 and 5.3 show a mirror situation. In fact, all the pairs in which the negation/antonymy have been correctly detected by EDITS$_{NEG}$ correspond to a negative behaviour of the system (i.e. the phenomenon generates contradiction), while all the pairs where lexical phenomena are detected by EDITS$_{LEX}$ correspond to a positive behaviour of the component.

As said before, beside the learned model, each TE-component outputs also a file with the sequence of edit operations that have been applied on T-H pairs, allowing us to carry out an error analysis on the data. Most of the wrong classifications are caused by situations of syntactic misalignment of constituents in T and H. Even if the use of the tree edit distance algorithm on the dependency trees of T and H should help us to deal with such cases, at the moment we are not able to fully exploit the advantages of this kind of representation. Once the dependency trees are created by the parser, the algorithm applies on them without considering the correctness of the trees. Furthermore, the algorithm implemented in the current version of EDITS

does not allow to perform the edit operations on subtrees or phrases.

More specifically, most of the mistakes of EDITS$_{COREF}$ are actually due to previous mistakes of the XIP coreference module, meaning that the terms were actually co-referent, but they were not recognized as such, so a high cost of substituting them is set. With respect to the negative behaviour of this component, we meant the pairs where the coreference between two entities is the cause of the unknown judgement in the pair (we are not sure if the two terms co-refer, e.g. *Mr Bouton* =? ⇒ *Daniel Bouton*). These cases are very rare in the data set, and EDITS$_{COREF}$ shows some problems in recognizing them.

Even if being able to correctly handle the phenomena covered by EDITS$_{NEG}$ is important to detect contradiction, their frequency in the data set is very low (only in 15 pairs out of 600 they are relevant to contradiction).[3] Most of EDITS$_{NEG}$'s mistakes in detecting negation (FN=26, Table 5.1) are due to the scarce coverage of the contradiction rules extracted from WordNet (i.e. in 53% of the cases). For example, in Example 5.5 (pair 298, RTE-5 training set) EDITS$_{COREF}$ correctly substitutes both the co-referent Named Entities in T and H, while EDITS$_{NEG}$ does not apply because there is no contradiction rules for the antonyms "opponent" - "ally".

(5.5) T: *The current Prime Minister Stephen Harper supported Mulroney's right to comment on Trudeau, "I think it's well known Mr. Mulroney was an opponent of Mr. Trudeau" Harper said. [...]*

H: *Mr Mulroney was an ally of Mr Trudeau.*

Other mistakes of EDITS$_{NEG}$ are due to the fact that the component applies in a pair where a negation is present, but such negation is meaningless to state if there is/ there is not contradiction in the pair - i.e. it shows a

[3]This aspect will be discussed in more details in Chapter 8, where we analyse contradiction pairs.

negative behaviour, while it should have shown a neutral one (FP=12, Table 5.1), as discussed also in Cabrio *et al.* 2008 [18].

EDITS$_{LEX}$ shows bad results in particular with respect to the neutral behaviour, because often the lexical substitution is carried out, but the wrong sub-sentence of T is chosen.[4] For instance, in Example 5.6 (pair 152 RTE-5 test set):

(5.6) T: *MANILA, Philippines - Fishermen in the Philippines accidentally caught and later ate a megamouth shark, one of the rarest fishes in the world [...]. The 1,100-pound, 13-foot-long* ***megamouth*** *died while struggling in the fishermen's net on March 30 off Burias island in the central Philippines.[...]*

H: *A megamouth is a rare species of* ***shark.***

"megamouth" is substituted with "shark"[5], but the sentence of T carrying the entailing meaning was the first one, and not the one chosen by the algorithm. Moreover, in other cases EDITS$_{LEX}$ substitutes at a low cost two words that are highly related according to the entailment rules, but that in that specific pairs should have not been substituted because of different reasons: *i)* words not related in that context (we will discuss this point in Chapter 7, where we propose a methodology to automatically acquire rules enriched with the context, to maximize precision), *ii)* semantically similar modifiers modifying different heads, *iii)* semantically related words but not replaceable (e.g. *mother* and *sister*) - this is due to the fact that we extracted rules from Wikipedia, so the coverage is broader with respect to WordNet, but the accuracy is lower.

[4]Wrong with respect to the one that should have been chosen in order to correctly assign the entailment judgement.

[5]Even if actually the word "megamouth" is present also in H in the same position occupied in T, so the algorithm should have chosen that substitution operation.

5.5 Combining TE-components

In Section 5.4.1 we described the TE-components we implemented basing on EDITS architecture, and in Section 5.4 we run them independently on RTE-5 to check if the expected behaviours we hypothesized are correctly put in practice. In this Section, we experiment two compositional models for the combination of the TE-components within EDITS architecture, namely the weighted linear composition and a sequential composition of the distances produced by the single modules on T-H pairs (Section 5.5.1). These models implement the inferential structures of the arguments we described in Chapter 4, i.e. the weighted linear composition reflects a convergent inferential structure, while the sequential composition reflects a non convergent one. With respect to the compositional mechanisms based on Natural Logic semantic relations we described in Chapter 4, the strategies we propose here correspond to preliminary steps, to verify the feasibility of the approach. Experimental results on RTE data demonstrate that the second model, that takes into account the dependencies among the linguistic phenomena, is superior when compared to the first one, suggesting that this is a promising direction to explore.

Since the considered phenomena are situated on different linguistic levels (e.g. lexical, syntactic and semantic), the distance provided by each module could impact in a different way on the general entailment judgement of a T-H pair, depending on the importance and on the granularity of the phenomenon it deals with. To take advantage of this intuition, the contribution of each component has been weighted, and such weights have been automatically learnt using Particle Swarm Optimization (PSO) methods, as allowed by EDITS (Mehdad 2009 [61]). Interestingly, as we will show in the experimental section (Section 5.5.2), the application of these methods brings to an improvement in results, and, from a more theoreti-

cal standpoint, makes the contribution of the different phenomena in the assignment of the correct entailment judgement more evident.

5.5.1 Compositional strategies

Although there can be several strategies for the combination of the distances produced by each TE-component to produce a unique result, and to correctly assign the entailment judgement, we experimented two compositional models: the weighted linear composition and a sequential composition of the distances produced by the single modules on T-H pairs.

Weighted Linear Composition. The first method for combining the TE-components is based on standard approaches for the combination of classification models (Kittler *et al.* 1998 [46]). According to this strategy, each module resolves the phenomenon it is build to address, and outputs the edit distance. Given a distance estimated by each module, the overall score can be derived as a linear combination (e.g. summation).

In order to estimate the importance and confidence of each module in the overall performance, the weighted linear combination is recommended. The weights are obtained by optimizing the performance on the training data using PSO methods. The final score is computed as:

$$ED(T,H) = \sum_{i=0}^{n} w_i D_i(T,H)$$

where $D_i(T,H)$ and w_i are the distance and the weight of the TE-component i, respectively.

Sequential Composition. This method runs the modules in a sequential order, and the output operations of each TE-component are considered as input of the next module. Basing on linguistic intuitions, components deal-

ing with phenomena that can change the polarity of a sentence will come first, since they can be the cause of contradiction in the pair. Furthermore, modules whose transformations apply to smaller portions of text are run first than the ones involving syntactic constructions.

The main difference of this composition strategy with respect to the linear combination, is that until the first component has not finished the job, the next one cannot interfere. Compared with the linear combination method, sequential combination has lower computational complexity, however it is more expensive in terms of time and implementation. Each module can be optimized while it runs.

5.5.2 Experiments and results

Aim of these preliminary experiments is to investigate which composition method among the ones proposed can effectively re-combine the component-based approach. In more detail, we set up two different sets of experiments using the EDITS-based TE-components we implemented - namely, EDITS$_{COREF}$, EDITS$_{NEG}$, and EDITS$_{LEX}$ (Section 5.3) - again on RTE-5 data (Table 5.4 presents the results).

To experiment the Weighted Linear Composition, the three TE-components were run in parallel, we summed the distances produced by each component, and then we learnt a model using SVM-light. Such model is then applied on the test set. On the contrary, to experiment the Sequential Composition, we run EDITS$_{NEG}$ first, since we want to detect negative polarity items that could be the cause of contradiction among the two fragments. Then, EDITS$_{COREF}$ is run to solve coreferences, followed by EDITS$_{LEX}$. We also tried to take advantage of a linguistic intuition on the dependencies among phenomena, setting for instance high costs of substituting two synonyms if one of them is negated, or setting low costs for substitutions among antonyms if one of them is negated. Each TE-component

was optimized and tuned using the method introduced in (Mehdad and Magnini 2009 [62]): the values reported in the right columns of Table 5.4 refer to the weights attributed to the modules in the linear model, and to the edit operations in the sequential one. As can be noticed, in both of them EDITS$_{NEG}$ has the highest weights, despite the low precision we obtained in the experiments presented in Section 5.4.

	Modules	**Weights**		
	All	**EDITS**$_{NEG}$	**EDITS**$_{COREF}$	**EDITS**$_{LEX}$
Linear	56.82%	0.8	0.4	0.1
Sequential	60.0%	1.0	0.1	0.2

Table 5.4: Results comparison over RTE5 data set.

The results show that sequential combination improves the linear method on the test set (about 3% in accuracy). At this point of the experimental phase, we cannot significantly compare our results with the performances of the TE systems submitted to previous RTE evaluation campaigns, since the coverage of the components we implemented is not high enough to draw final conclusions. However, the benefits of the idea underlying the component-based framework are shown and experimented through a comparison of different composition strategies. We expect that by developing more precise, and a higher number of TE-components to augment the coverage of the phenomena considered would improve our results. In particular, more phenomena that can cause contradiction in the pair (e.g. quantity or temporal expression mismatching) should be faced.

5.6 Participation at RTE evaluation campaigns

A preliminary version of the component-based architecture based on EDITS we described in this Chapter has been evaluated in our participations

to RTE campaigns, in particular in RTE-4 (Cabrio *et al.* 2008 [18]). Two TE-components, namely EDITS$_{NEG}$ to deal with negative polarity items, and EDITS$_{LEX}$ to deal with lexical similarity were part of the architecture, and the Linear Distance algorithm was used. More specifically, EDITS$_{LEX}$ was not set as an independent component, but was integrated in a more general module that considered all but the negation phenomena, plus WordNet similarities (EDITS$_{ALL-BUT-NEG}$). Entailment rules exploited by this module were extracted from WordNet basing on the relation of synonymy, and basing on WordNet similarity package (the Adapted Lesk - Extended Gloss Overlaps measure, Pedersen *et al.* 2004 [78]). Our official results at RTE-4 Challenge are shown in Table 5.5. We submitted three runs for the two-way RTE task: the first one with EDITS$_{NEG}$, the second one with a combined system (EDITS$_{NEG}$+ EDITS$_{ALL-BUT-NEG}$) and the third one with a standard configuration of the EDITS system.

	first run	**second run**	**third run**
accuracy %	54	54.6	57
avg. precision %	49.4	55.1	55.3

Table 5.5: Results on RTE-4 data set

Concerning the first two runs, we participated in the RTE challenge as a way to understand what our modular system could do with respect to more general systems used in RTE. Given the promising results, we were encouraged to continue with this research line.

In our participation at RTE-5 and 6, we mainly submitted system runs using standard configurations of the EDITS system tuned for the challenges, in order to experiment both different knowledge resources (in RTE-5, Mehdad *et al.* 2009 [64]), and different algorithms (in RTE-6, Kouylekov *et al.* 2010 [48]). However, in both challenges the TE-component EDITS$_{COREF}$ was used to detect the co-referent terms and to assign specific

costs to this operation.

5.7 Conclusions and future work

Basing on the theoretical definitions of the TE component-based architecture we proposed in Chapter 4, in this Chapter we carried out some experiments to prove the feasibility of the described approach. We took advantage of the flexible and modular architecture of the EDITS system, and we implemented a set of TE-components, compliant with the criteria we previously defined. We first independently ran each TE-component on RTE-5 data to check if the expected behaviours were put in practice, and then we experimented two different strategies to combine the output produced by each component to obtain an overall judgement for a pair. The experimental setup presents some simplifications with respect to the combination strategy based on an extended model of Natural Logic we presented in Chapter 4, and the results we obtained are not completely satisfying. However, starting from these preliminary experiments we plan to refine the design of the TE-components to improve the single precisions with respect to the expected behaviours, also considering an algorithm different from the edit distance algorithm, that turned out not to be the optimal one. Furthermore, the number of the implemented components should be augmented, in order to broaden the coverage of the considered linguistic phenomena. This way, we expect to obtain a general improvement in the performances on RTE data. Increasing the number of components will bring up even with more evidence the sequential order issue, that is a very interesting direction to explore, taking advantage of the dependencies among the linguistic phenomena in the data, as discussed in Chapter 4.

Chapter 6

Textual Entailment Specialized Data Sets

This Chapter presents the pilot study we carried out for the creation of specialized data sets for TE, made of atomic T-H pairs, i.e. pairs in which a certain phenomenon relevant to the entailment relation is highlighted and isolated (Bentivogli et al. 2010 [11]). The result is a resource that can be profitably used both to advance in the comprehension of the linguistic phenomena relevant to entailment judgements, and to make a first step towards the creation of large-scale specialized data sets.

6.1 Introduction

In Chapter 4 we pointed out that to correctly judge each single pair inside the RTE data sets, systems are expected to cope both with the different linguistic phenomena involved in TE, and with the complex way in which they interact. But one of the major issues raised by the TE community is that while system developers create new modules, algorithms and resources to address specific inference types, it is difficult to measure a substantial impact when such modules are evaluated on the RTE data sets because of *i)* the sparseness (i.e. low frequency) of the single phenomena, and

ii) the impossibility to isolate each phenomenon, and to evaluate each module independently from the others. Recently, Sammons *et al.* (2010) [83] sought to start a community-wide effort to annotate RTE examples with the inference steps required to reach a decision about the example label (entailment vs. contradiction vs. unknown)[1]. The authors propose a linguistically-motivated analysis of entailment data based on a step-wise procedure to resolve entailment decision, by first identifying parts of T that match parts of H, and then identifying connecting structures (see Chapter 3). This work is very similar in spirit to the approach we propose here, and shows the interest of the TE community towards this research direction.

Basing on the methodology for the creation of atomic T-H pairs we described and motivated in Chapter 4, we propose to cluster all the atomic pairs related to a certain phenomenon to create specialized TE data sets, to allow systems training and evaluation. Summing up briefly, the proposed methodology starts from an existing RTE pair and defines the following steps: *i)* identify the phenomena present in the original RTE pair; *ii)* apply an annotation procedure to isolate each phenomenon and create the related atomic pair; finally, *iii)* group together all the atomic T-H pairs relative to the same phenomenon, hence creating specialized data sets. The expected benefits of specialized data sets for TE derive from the intuition that investigating the linguistic phenomena separately, i.e. decomposing the complexity of the TE problem, would yield an improvement in the development of specific strategies to cope with them. In fact, being able to detect entailment basing on linguistic foundations should strengthen the systems, making the overall performances less data set dependent. We carried out a feasibility study applying the devised methodology to a sample of 90 pairs extracted from the RTE-5 data set (Bentivogli *et al.* 2009 [14])

[1] `https://agora.cs.illinois.edu/display/rtedata/Explanation+Based+Analysis+of+RTE+Data`

and we addressed a number of critical issues, including: *i)* whether it is possible to clearly identify and isolate the linguistic phenomena underlying the entailment relation; *ii)* how specific the categorization of phenomena should be; *iii)* how easy/difficult it is to create balanced data sets of atomic T-H pairs with respect to the distribution of positive and negative examples, so that these data sets might be used for training and testing. In Section 6.2 we describe the annotation procedure for the creation of the specialized data sets, based on the procedure to create the atomic pairs described in Chapter 4. In Section 6.3 some examples of the application of the methodology are presented, while in Section 6.4 a feasibility study carried out on a sample of the RTE-5 data set is described and the resulting data are given. The result of the feasibility study is a *pilot* resource, freely available for research purposes.[2] In Section 6.5 a number of issues that arise while trying to create a balanced data set are presented; Section 6.6 draws some final remarks and discusses on the feasibility of the proposed approach for the creation of large-scale data sets.

6.2 Methodology for the creation of atomic T-H pairs

In this Section we recap the methodology defined in Chapter 4, with the aim of applying it systematically to RTE data sets. The idea is to create atomic pairs[3] on the basis of the phenomena which are actually present in the RTE T-H pairs. One of the advantages of applying the methodology to the RTE data consists of the fact that the actual distribution of the linguistic phenomena involved in the entailment relation emerges. In Chapter 4 we proposed a classification of the phenomena we detected while analysing a sample of RTE pairs, and we decided to group them us-

[2] `http://hlt.fbk.eu/en/Technology/TE_Specialized_Data`

[3] In our previous papers, we used to refer to the atomic T-H pairs as *monothematic* pairs. In this Thesis we decided to switch the terminology to be compliant with the theoretical framework we propose.

ing both fine-grained categories and broader categories. Grouping specific phenomena into macro categories allows us to create specialized data sets containing enough pairs to train and test TE systems. Macro categories are defined referring to widely accepted linguistic categories in the literature (e.g. Garoufi 2007 [38]) and to the inference types typically addressed in RTE systems: lexical, syntactic, lexical-syntactic, discourse and reasoning. In Chapter 4 we defined the notion of entailment rule, as a formalization of the knowledge about a linguistic phenomenon relevant to TE.

Given such basic concepts, the procedure consists of a number of steps carried out manually. We start from a T-H pair taken from one of the RTE data sets and we decompose T-H in a number of atomic pairs T-H_i, where T is the original Text and H_i are Hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in T-H. The procedure is schematized in the following steps:

1. Individuate the linguistic phenomena which contribute to the entailment in T-H

2. For each phenomenon i:

   (a) individuate a general entailment rule r_i for the phenomenon i, and instantiate the rule using the portion of T which expresses i as the LHS of the rule, and information from H on i as the RHS of the rule.

   (b) substitute the portion of T that matches the LHS of r_i with the RHS of r_i.

   (c) consider the result of the previous step as H_i, and compose the atomic pair $T - H_i$. Mark the pair with phenomenon i.

3. Assign an entailment judgement to each atomic pair.

After applying this procedure to the original pairs, all the atomic $T-H_i$ pairs relative to the same phenomenon i should be grouped together in a data set specialized for phenomenon i.

6.3 Application of the procedure to RTE pairs

In this section, we show examples of the application of the procedure to RTE pairs, namely entailment (Section 6.3.1), contradiction (Section 6.3.2) and unknowns pairs (Section 6.3.3).

6.3.1 Entailment pairs

Table 6.1 shows the decomposition of an original entailment pair (pair 199 in RTE-5) into atomic pairs. At step 1 of the methodology, the phenomena (i.e. modifier, coreference, transparent head and general inference) are considered relevant to the entailment between T and H. In the following, we apply step by step the procedure to the phenomenon we define as modifier. At step 2a the general rule:

Entailment rule:	**modifier**
Pattern:	$X\ Y \Leftrightarrow\ Y$
Constraint:	*MODIFIER(X,Y)*
Probability:	1

is instantiated (*The tiny Swiss canton* $\Rightarrow$ *The Swiss canton*), while at step 2b the substitution in T is carried out (*The Swiss canton of Appenzell Innerrhoden has voted to prohibit [...]*[4]).

At step 2c the atomic pair $T-H_1$ is composed and marked as *modifier* (macro-category *syntactic*). Finally, at step 3, this pair is judged as *entail-*

[4]The symbol [...] is used as a placeholder of the missing parts.

Text snippet (pair 199 RTE-5 test set)			**Rule**	**Phenomena**	**Judg.**
T		**The tiny Swiss canton of Appenzell Innerrhoden has voted to prohibit the phenomenon of naked hiking.** Anyone found wandering the Alps wearing nothing but a sturdy pair of hiking boots will now be fined.			
H		The Swiss canton of Appenzell has prohibited naked hiking.		**synt:modifier, disc:coref, lexsynt:tr_head, reas:gen_infer**	**E**
	H_1	**The Swiss canton** of Appenzell Innerrhoden has voted to prohibit the phenomenon of naked hiking.	**x y ⇒ y modif(x,y)**	**synt:modifier**	**E**
	H_2	The tiny Swiss canton of **Appenzell** has voted to prohibit the phenomenon of naked hiking.	**x⇔y coref(x,y)**	**disc:coref**	**E**
	H_3	The tiny Swiss canton of Appenzell Innerrhoden has voted to prohibit **naked hiking**.	**x of y ⇒y tr_head(x,y)**	**lexsynt:tr_head**	**E**
	H_4	The tiny Swiss canton of Appenzell Innerrhoden **prohibited** the phenomenon of naked hiking.	**vote to prohibit (+ will now be fined) ⇒ prohibit**	**reas:gen_infer**	**E**

Table 6.1: Application of the decomposition methodology to an *entailment* pair.

ment. Step 2 (a, b, c) is then repeated for all the phenomena individuated in that pair at step 1.

It can be the case that several phenomena are collapsed on the same token, as in Example 4.1 we showed in Chapter 4. In such cases, in order to create an atomic H for each phenomenon, the methodology is applied recursively. It means that after applying it once to the first phenomenon of the chain (therefore creating the pair $T - H_i$), it is applied again on H_i (that becomes T') to solve the second phenomenon of the chain (creating the pair $T' - H_j$).

6.3.2 Contradiction pairs

Table 6.2 shows the decomposition of an original contradiction pair (pair 125 in RTE-5) into atomic pairs. At step 1 both the phenomena that preserve the entailment and the phenomena that break the entailment rules causing a contradiction in the pair should be detected. In the example reported in Table 6.2, the phenomena that should be solved in order to correctly judge the pair are: argument realization, apposition and semantic opposition. While the atomic pairs created basing on the first two phenomena preserve the entailment, the semantic opposition generates a contradiction. In the following, we apply step by step the procedure to the phenomenon of semantic opposition (Chapter 4).
At step 2a the general rule:

Contradiction rule:	**semantic_opposition**
Pattern:	$X \nLeftrightarrow Y$
Constraint:	*SEMANTIC_OPPOSITION(Y,X)*
Probability:	1

is instantiated (*new* $\nLeftrightarrow$ *outgoing*), and at step 2b the substitution in T is carried out (*Mexico's outgoing president, Felipe Calderon [...]*). At step

	Text snippet (pair 408 RTE-5 test set)		Rule	Phenomena	Judg.
T	**Mexico's new president, Felipe Calderon,** seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]				**C**
H	Felipe Calderon is the outgoing President of Mexico.			**lex:sem_opp synt:arg_realiz synt:apposit**	
	H_1	Mexico's **outgoing** president, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]	$x \nLeftrightarrow$ **y**	**sem_opp(x,y)**	**C**
	H_2	The new president **of Mexico**, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]	**x's y ⇒ y of x**	**synt:arg_realiz**	**E**
	H_3	**Felipe Calderon is** Mexico's **new president**.	**x,y ⇒ y is x apposit(y,x)**	**synt:apposit**	**E**

Table 6.2: Application of the decomposition methodology to a *contradiction* pair.

2c a negative atomic pair $T - H_1$ is composed and marked as semantic opposition (macro-category *lexical*), and the pair is judged as *contradiction*. We noticed that negative atomic T-H pairs (i.e. both contradiction and unknown) may originate either from the application of contradiction rules (e.g. semantic opposition or negation, as in pair $T - H_1$, in Table 6.2) or as a wrong instantiation of a positive entailment rule. For instance, the positive rule for active/passive alternation:

Entailment rule:	**active/passive_alternation**
Pattern:	$X\ Y\ Z \Leftrightarrow Z\ W\ X$
Constraint:	*SAME_STEM(X,W)*
	TYPE(X)=V_ACT; *TYPE(W)=V_PASS*
Probability:	1

when wrongly instantiated, as in *Russell Dunham killed nine German soldiers* ⇏ *Russell Dunham was killed by nine German soldiers* ($x\ y\ z \Leftrightarrow z\ w\ x$), generates a negative atomic pair.

6.3.3 Unknown pairs

Table 6.3 shows the decomposition of an original unknown pair (pair 82 in RTE-5) into atomic pairs. At step 1 all the relevant phenomena are detected: coreference, general inference, and modifier.

Text snippet (pair 82 RTE-5 test set)				Rule	Phenomena	Judg.
T	Currently, there is no specific treatment available against **dengue fever, which is the most widespread tropical disease after malaria.** [...] "Controlling **the mosquitos that transmit dengue** is necessary but not sufficient to fight against the disease [...]"					
H	Malaria is the most widespread disease transmitted by mosquitos.				**disc:coref, reas:gen_infer, synt:modifier,**	**U**
	H_1 $\rightarrow T'$	**Dengue fever** is the most widespread tropical disease after malaria.		**x⇔y coref(x,y)**	**disc:coref**	**E**
		H_2	**Malaria is the most widespread** tropical disease.	**x is after y⇒ y is the first**	**reas:gen_infer**	**E**
		H_3	Dengue fever is the most widespread disease **trasmitted by mosquitos** after malaria.	**x =? ⇒ x y (restr. relat. clause)**	**synt:modifier**	**U**

Table 6.3: Application of the methodology to an *unknown* pair.

While the first two preserve the entailment relation, the atomic pair

resulting from the third phenomenon is judged as unknown. As discussed in Chapter 4, the last atomic pair is an argument with a very low inductive probability (i.e. the fact that a certain disease is the most widespread among the ones transmitted by a certain cause, does not allow us to infer that it is the most widespread ever). If we try to apply step by step the procedure to the phenomenon of modifier, at step 2a the generic rule:

Entailment rule:	**modifier**
Pattern:	$X \Rightarrow X\ Y$
Constraint:	*MODIFIER(Y,X)*
Probability:	0.1

is instantiated (*disease* $\Rightarrow$ *disease transmitted by mosquitoes*) (this rule has a very low probability), and at step 2b the substitution in T is carried out. At step 2c the atomic pair T'-H_3 is composed and marked as *modifier* (restrictive relative clause, macro-category *lexical*), and the pair is judged as *unknown*. However, as already stated in Chapter 4, there is no reason to collect such kind of rules for computational purposes, since it would mean to collect almost all the relations among all the words and the expressions of a language. These rules are somehow obtained in a complementary way with respect to high-probability rules, i.e. if a certain rule is not present among the highly probable ones, it means that it has a low probability, and therefore it is not strong enough to support the related inferential step.

6.4 Feasibility study on RTE-5 data

In order to assess the feasibility of the specialized data sets, we applied our methodology to a sample of 90 T-H pairs randomly extracted from the RTE-5 data set. In particular, the sample pairs are equally taken from

entailment, *contradiction* and *unknown* examples.

6.4.1 Inter-annotator agreement

The whole RTE-5 sample has been annotated by two annotators with skills in linguistics and inter-annotator agreement has been calculated. A first measure of *complete* agreement was considered, counting when judges agree on all phenomena present in a given original T-H pair. The complete agreement on the full sample amounts to 64.4% (58 up to 90 pairs). In order to account for partial agreement on the set of phenomena present in the T-H-pairs, we used the *Dice coefficient* (Dice 1945 [34]).[5] The Dice coefficient is computed as follows:

$$Dice = 2C/(A + B)$$

where C is the number of common phenomena chosen by the annotators, while A and B are respectively the number of phenomena detected by the first and the second annotator. Inter-annotator agreement on the whole sample amounts to 0.78. Overall, we consider this value high enough to demonstrate the stability of the (micro and macro) phenomena categories, thus validating their classification model. Table 6.4 shows inter-annotator agreement rates grouped according to the type of the original pairs, i.e. *entailment*, *contradiction* and *unknown* pairs.

The highest percentage of *complete* agreement is obtained on *unknown* pairs. This is due to the fact that since the H in *unknown* pairs typically contains information which is not present in (or inferable from) T, for 19

[5]The *Dice coefficient* is a typical measure used to compare sets in IR and is also used to calculate inter-annotator agreement in a number of tasks where an assessor is allowed to select a set of labels to apply to each observation. In fact, in these cases, and in ours as well, measures such as the widely used K are not good to calculate agreement. This is because K only offers a dichotomous distinction between agreement and disagreement, whereas what is needed is a coefficient that also allows for partial disagreement between judgements.

pairs out of 30 both the annotators agreed that no linguistic phenomena relating T to H could be detected.

	Complete	**Partial (Dice)**
ENTAILMENT	60%	0.86
CONTRADICTION	57%	0.75
UNKNOWN	76%	0.68

Table 6.4: Agreement measures per entailment type

With respect to the Dice coefficient, the highest inter-annotator agreement can be seen for the *entailment* pairs, whereas the agreement rates are lower for *contradiction* and *unknown* pairs. This is due to the fact that for the *entailment* pairs, all the single phenomena are directly involved in the entailment relation, making their detection straightforward. On the contrary (cfr. Sections 6.3.2 and 6.3.3), in the original *contradiction* and *unknown* pairs not only the phenomena directly involved in the contradiction/unknown relation are to be detected, but also those preserving the entailment, which do not play a direct role on the relation under consideration (contradiction/unknown) and are thus more difficult to identify.

6.4.2 Results of the feasibility study

The distribution of the phenomena present in the original RTE-5 pairs, as resulting after a reconciliation phase carried out by the annotators, is shown in Table 6.5. The total number of occurrences of each specific phenomenon is given (Column *TOT*), corresponding to the number of atomic pairs created for that phenomenon. The number of atomic pairs is then broken down into positive examples - i.e. *entailment* atomic pairs (Column *E*) - and negative examples - i.e. *contradiction* and *unknown* atomic pairs (Columns *C* and *U*, respectively).

A number of remarks can be made on the data presented in Table 6.5. Both macro categories and fine-grained phenomena are well represented but show a different absolute frequency: some have a high number of occurrences, whereas some others occur very rarely. In particular, as already pointed out in Garoufi (2007) [38], also our study confirms that the phenomena belonging to the category *reasoning* are the most frequent, meaning that a significant part of the data involves deeper inferences.

As for the distribution among E/C/U atomic pairs, we can see that some phenomena appear more frequently - or only - among the positive examples (e.g. apposition or coreference) and others among the negative ones (e.g. quantitative reasoning). In general, the total number of positive examples is much higher than that of the negative ones and, for some macro-categories (e.g. lexical-syntactic) no negative examples are found. Also from a qualitative standpoint, the variability of phenomena in negative examples is reduced with respect to the positive pairs.

Overall, the feasibility study showed that the decomposition methodology we propose can be applied on RTE-5 data. The task demonstrated to be feasible under a number of aspects. As for the quality of the atomic pairs, the high inter-annotator agreement rate obtained shows that the methodology is stable enough to be applied on a large scale. With respect to the human effort required, during the feasibility study an average of four original RTE-5 pairs per hour have been decomposed. This means that, provided that the task be carried out by annotators with a curriculum in linguistics, around two and a half person months are required to apply the decomposition methodology to the whole RTE-5 data set, which is composed of 1200 T-H pairs.

Phenomena	**Atomic Pairs**			
	TOT	**E**	**C**	**U**
Lexical:	**32**	**22**	**8**	**2**
Identity/mismatch	4	1	3	0
Format	2	2	0	0
Acronymy	3	3	0	0
Demonymy	1	1	0	0
Synonymy	11	11	0	0
Semantic opposition	3	0	3	0
Hypernymy	5	3	0	2
Geographical knowledge	3	1	2	0
Lexical-syntactic:	**18**	**18**	**0**	**0**
Transparent head	3	3	0	0
Nominalization/verbalization	9	9	0	0
Causative	1	1	0	0
Paraphrase	5	5	0	0
Syntactic:	**44**	**30**	**10**	**4**
Negation	1	0	1	0
Modifier	3	3	0	0
Argument Realization	6	6	0	0
Apposition	17	11	6	0
List	1	1	0	0
Coordination	5	4	0	2
Active/Passive alternation	6	4	2	0
Discourse:	**44**	**43**	**0**	**1**
Coreference	24	23	0	1
Apposition	3	3	0	0
Anaphora Zero	12	12	0	0
Ellipsis	4	4	0	0
Statements	1	1	0	0
Reasoning:	**67**	**45**	**17**	**6**
Apposition	3	2	1	0
Modifier	3	3	0	0
Genitive	1	2	0	0
Relative Clause	1	1	0	0

Elliptic Expression	1	1	0	0
Meronymy	4	3	1	0
Metonymy	3	3	0	0
Membership/representative	2	2	0	0
Quantity	6	0	5	1
Temporal	2	1	0	1
Spatial	1	1	0	0
Common background/ general inferences	40	26	20	4
TOTAL (# atomic pairs)	**206**	**158**	**35**	**13**

Table 6.5: Distribution of phenomena in T-H pairs.

6.5 Creating Specialized Data sets

After applying the procedure described in Chapter 4 to the original 90 pairs of our sample, all the atomic $T-H_i$ pairs relative to the same phenomenon i can be grouped together, resulting in several data sets specialized for phenomenon i. For instance, we can create a specialized data set for Reasoning phenomena, which would include 67 atomic pairs, out of which 45 are positive, 17 are contradiction and 6 are unknown (see Table 6.5).

As introduced before, due to the natural distribution of phenomena in RTE data, we found out that applying the decomposition methodology we generate a higher number of atomic positive pairs (76.7%) than negative ones (23.3%, divided into 17% *contradiction* and 6.3% *unknown*, as shown in Table 6.5). We analysed separately the three subsets composing the RTE-5 sample, (i.e. 30 *entailment* pairs, 30 *contradiction* pairs, and 30 *unknown*) in order to verify the productivity of each subset with respect to the atomic pairs created from them. Table 6.6 shows the absolute distribution of the atomic pairs among the three RTE-5 classes.

When the methodology is applied to RTE-5 *entailment* examples, av-

		Phenomena / atomic pairs			
RTE-5 pairs		**E**	**C**	**U**	**Total**
	E (30)	91	–	–	91/30
	C (30)	44	35	–	79/30
	U (30)	23	–	13	36/11

Table 6.6: Distribution of the atomic pairs with respect to original E/C/U pairs

eragely 3.03 all positive atomic pairs are derived. When the methodology is applied to RTE-5 *contradiction* examples, we can create an average of 2.64 atomic pairs, among which 1.47 are entailment pairs and 1.17 are contradiction pairs. This means that the methodology is productive for both positive and negative examples.

As introduced before, in 19 out of 30 *unknown* examples no atomic pairs can be created, due to the lack of specific phenomena relating T and H (typically the H contains information which is neither present in T nor inferable from it). For the 11 pairs that have been decomposed into atomic pairs, we created an average of 3.27 atomic pairs, among which 2.09 are entailment and 1.18 are unknown pairs. This analysis shows that the only source of negative atomic pairs are the *contradiction* pairs, which actually correspond to 15% of RTE-5 data set.

As regards the issue of balancing each single specialized data set with respect to positive and negative examples (i.e. finding a balanced number of positive and negative examples for each single phenomenon) we saw in Section 6.4 that some phenomena appear more frequently - when not only - among the positive examples (e.g. apposition or coreference) while others appear more among the negative ones (e.g. quantitative reasoning). It happens that not only for specific phenomena but also for entire macro categories (e.g. lexical-syntactic) negative examples cannot be found. Al-

though the specialized data sets derived from the decomposition procedure might be useful for interesting corpus analysis investigations, current systems based on machine learning approaches would benefit from data sets with a more balanced proportion of negative examples. To cope with this problem, we devised a tentative solution, which consists of taking a positive example for a given phenomenon and synthetically creating a corresponding negative example by modifying the entailment rule. Starting from the observation of original *contradiction* and *unknown* pairs described in Section 6.3.2 and 6.3.3, we spotted out some possible operations to invalidate the rule which preserves the entailment in positive examples:

- invert a directional rule

  Pair 187, RTE-5 (phenomenon: REASONING:MODIFIER):

  T: [...] Islands are mostly made up of mangrove trees.

  H_1-pos: Mangroves are a kind of tree.

  H_1-neg: Trees are a kind of mangrove.

- wrongly instantiate a rule

  Pair 408, RTE-5 (phenomenon: LEXICAL:VERBALIZATION):

  T: [...] Doris Lessing, recipient of the 2007 Nobel Prize [...]

  H_3-pos: Doris Lessing received the 2007 Nobel Prize.

  H_3-neg: Doris Lessing receipted the 2007 Nobel Prize

  In this example the verbalization rule is wrongly instantiated by using a verb with the same stem of the verb "receive" but with another meaning.

- where possible, substitute the rule with another rule related to an opposite phenomenon.

  Pair 408, RTE-5 (phenomenon: LEXICAL:SYNONYMY):

T': [...] Doris Lessing received the 2007 Nobel Prize [...]

H_4-pos: Doris Lessing won the 2007 Nobel Prize.

H_4-neg: Doris Lessing refused the 2007 Nobel Prize.

This operation exploits the natural opposition of some phenomena (e.g. identity vs. negation; synonymy vs. oppositeness). In the example, the verb "win", which is synonym of "receive" is substituted with the verb "refuse", which is semantically opposed to "receive".

Two annotators carried out a study on the RTE-5 sample and found out that it was a difficult and time-consuming task leading to low inter-annotator agreement. For this reason, we suggest that alternative strategies for the generation of negative atomic pairs be further discussed. How to collect more negative examples is still an open issue, that deserves further investigation.

6.6 Conclusions

In this Chapter we based on the methodology described in Chapter 4 for the creation of specialized TE data sets, made of atomic T-H pairs in which a certain phenomenon underlying the entailment relation is highlighted and isolated. We carried out a pilot study applying such methodology to a sample of 90 pairs extracted from the RTE-5 data set and we demonstrated the feasibility of the task, both in terms of quality of the new pairs created and of time and effort required. An important outcome of the methodology proposed is that we provide the annotation of previous RTE data with the linguistic phenomena underlying the entailment/contradiction relations in the pairs (both with fine grained and macro categories), highlighting their actual distribution in the data, and allowing evaluations of the TE systems on specific phenomena both when isolated and when interacting with the

others. The result of our study is a new resource that can be used for training TE systems on specific linguistic phenomena relevant to inference.

Basing on the outcome and the considerations arisen in this pilot study, in Chapter 7 we experiment a strategy to automatically extract atomic pairs and entailment rules from Wikipedia revision history, with the goal of creating large-scale specialized data sets.

Chapter 7

Automatic Acquisition of Entailment Rules for Atomic T-H pairs

In this Chapter we propose a methodology for the automatic acquisition of atomic T-H pairs and, in particular, of the entailment rules that allow to carry out the related inferential step. We take advantage of the syntactic structure of atomic pairs to define the more appropriate linguistic constraints for the rule to be successfully applicable. We have carried out a large-scale application of our methodology on Wikipedia versions.

7.1 Introduction

In Chapter 4 we have introduced the notion of entailment rule (defined by Szpektor *et al.* 2007 [86]), as a directional relation between two sides of a pattern, corresponding to text fragments with variables (typically phrases or parse sub-trees, according to the granularity of the phenomenon they formalize). In our component-based framework, the linguistic knowledge expressed in the form of entailment rules provides the pieces of evidence needed to carry out a step of reasoning on a particular sub-problem of entailment present in a certain atomic T-H pair. More specifically, we define the allowed transformations (i.e. atomic edits) for a certain phenomenon

through a set of entailment rules for that specific phenomenon. As an example, given a T-H pair, a lexical rule like:

Entailment rule:	**synonymy_1**
Pattern:	*home* $\Leftrightarrow$ *habitation*
Probability:	0.8

expresses that the word *home* in Text can be aligned, or transformed, into the word *habitation* in the Hypothesis, with a probability equal to 0.8 that this operation preserves the entailment relation among T and H (as discussed in Chapter 4). Similar considerations apply for more complex rules, involving verbs, like:

Entailment rule:	**general_inference_1**
Pattern:	*X manufactures Y* $\Rightarrow$ *X's Y factory*
Probability:	0.8

where the variables may be instantiated by any textual element with a specified syntactic relation with the verb. Both kinds of rules are typically acquired either from structured sources (e.g. WordNet, Fellbaum 1998 [36]), or from semi-structured sources, like Wikipedia pages. As such sources do not provide an adequate representation of the linguistic context in which the rules can be successfully applied, their concrete use reflects this limitation. For instance, rule (1) (extracted from WordNet) would fail to be applied in a T-H pair where the sense of *home* is not a synonym of *habitation*, resulting in a decrease of the system's precision. The lack of linguistic knowledge constraints is also evident where knowledge is automatically extracted from unstructured sources according to distributional properties (e.g. DIRT (Lin and Pantel 2001 [51]). These rules

suffer from lack of directionality, and from low accuracy (i.e the strength of association of the two sides of the rule is often weak, and not well defined). For instance, in rule (2) (extracted from DIRT), no directionality is expressed, and additional constraints to specify the variables types are required to correctly instantiate them. These observations are also in line with the discussion on ablation tests carried out at the last RTE evaluation campaigns (Bentivogli *et al.* [14]).

According to the considerations above, we have addressed the acquisition of high-precision entailment rules under a novel perspective. We take advantage of material obtained through Wikipedia revisions, which provides at the same time real textual variations from which we may extrapolate relevant linguistic context, and several simplifications with respect to alternative resources. Specifically, we consider T-H pairs where T is a revision of a Wikipedia sentence and H is the original sentence, as the revision is considered more informative then the revised sentence. Starting from such T-H pairs we could optimize crucial aspects of the acquisition procedure, including:

- *Rule precision.* Wikipedia revisions typically involve few differences; consequently, it is relatively easy to isolate the portion of sentence which may originate an entailment rule. Under this perspective, Wikipedia T-H pairs are more suitable for rule extraction with respect to more complex RTE pairs. An additional factor is that the amount of Wikipedia revisions is huge (and constantly increasing), which means that we can exploit redundancy in order to improve confidence.

- *Rule directionality.* It has been observed (Zanzotto and Pennacchiotti 2010 [99]) that, in most of the cases, the revision of a Wikipedia sentence preserves the entailment relation with respect to the original sentence. This allows us to assume, at least with a reasonable approx-

imation, that rules derived from Wikipedia revision pairs maintain the same direction of Text and Hypothesis. The qualitative analysis of the resulting resources extracted from Wikipedia revision pairs has confirmed this assumption.

- *Rule linguistic context.* The fact that Wikipedia revision pairs show few differences, opens the possibility to isolate the specific phenomena relevant for entailment with an acceptable accuracy. The consequence is that we could detect the appropriate syntactic context of the rule, in terms of constraints such that maximize the successful application of the rule.

To show the feasibility of the acquisition of high precision rules from Wikipedia revision pairs, we have carried out two large-scale experiments focusing, respectively, on entailment rules for *causality* and *temporal expressions*. Both phenomena are highly frequent in Textual Entailment pairs (see Chapter 6), and for both there are no available resources yet. The result consists in a large repository (freely available for research purposes)[1] that can be used by Textual Entailment systems, and that can be easily extended to entailment rules for other phenomena.

7.2 Related Work

The interest of the research community in producing specific methods to collect inference and paraphrase pairs is proven by a number of different works in the field, which are relevant for the approach we propose in this Chapter. As for paraphrase, Sekine's Paraphrase Database (Sekine 2005 [84]) was collected using an unsupervised method, and focuses on phrases which connect two Named Entities. In the Microsoft Research Paraphrase

[1] `http://hlt.fbk.eu/en/technology`

Corpus[2], 5800 pairs of sentences have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship. Since they are paraphrase collections, in both data sets rules are bidirectional, while one of the peculiarity of the entailment relation is the directionality, which we address in our work.

As for rule repositories collected using distributional properties, DIRT (Discovery of Inference Rules from Text)[3] is a collection of inference rules (described in Chapter 3, Lin and Pantel 2001 [51]), obtained extracting paths (binary relations) from dependency trees. The slot fillers in the path are nouns because slots correspond to variables in inference rules and are instantiates by entities; internal relations are between a verb and an object-noun or a small clause. Also in this case rules are not directional.

More recently, Aharon *et al.* (2010) [80] presented FRED, an algorithm for generating entailment rules between predicates from FrameNet. Annotated sentences and relations between frames are used to extract both the entailment relations and their argument mappings.

Szpektor *et al.* (2004) [87] produce the TEASE collection of entailment rules, automatically acquired from the web. The current collection consists of 136 different templates that were given as input, plus all the learned templates for that input template. The algorithm for Web-based extraction of entailment relations is applied for acquiring entailment relations for verb-based expressions. Also TEASE does not specify the directionality of the produced template pairs, but additional mechanisms that attempt to guess the directionality have been proposed. A manually created rule base for generic linguistic phenomena, e.g. syntactic-based rules, (e.g. conjunctions, clausal modifiers, relative clauses, appositives) is described in

[2] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

[3] http://www.aclweb.org/aclwiki/index.php?title=DIRT_Paraphrase_Collection

(Bar-Haim *et al.* 2007 [7]). In this case the scope of the resource is limited to few specific phenomena.

The use of Wikipedia revision history in NLP tasks has been previously investigated by (Zanzotto *et al.* 2010 [99]) and (Max and Wisniewski 2010 [60]). In the first work, two versions of Wikipedia and semi-supervised machine learning methods are used to extract large textual entailment data sets similar to the ones provided for the RTE challenge. In (Max and Wisniewski 2010 [60]), the revision history of this resource is used to create a corpus of natural rewritings, that includes spelling corrections, reformulations, and other local text transformations.

As discussed in Chapter 3, because of its high coverage Wikipedia is used by some TE systems for extraction of lexical-semantic rules, Named Entity Recognition and geographical information. However, so far it has only been used as source of factual knowledge, while in this Chapter the focus is on the acquirement of more complex rules, concerning for instance spatial or temporal expressions.

7.3 General Methodology

The general approach we have implemented is based on the idea that, given a *seed word*, we want to extract, from Wikipedia revision pairs, all the entailment rules where the seed word appears as the head of the rule (i.e. the non-variable part of the rule from which the other parts depend on - for instance the word *manufactures* is the head of rule *general_inference_2* in Section 7.1), either in T or H. Wikipedia revision pairs, because of their specific nature, allow to simplify the rule extraction task in a number of aspects, which are discussed in this Section.

Entailment judgement. A Wikipedia revision may be consistent with

the original sentence, in which case it brings an entailment relation, or it may introduce inconsistency, in which case it expresses a contradiction relation with respect to the original sentence. For our experiments we have assumed that a great proportion (i.e. about 95%) of revisions preserves entailment, and that this is the default case. This assumption is in line with (Zanzotto and Pennacchiotti 2010 [99]), and has been confirmed by manually checking a sample of revision pairs.

Atomic T-H pairs. The capability of automatic extraction of entailment rules is affected by the complexity of the pairs from which we extract the rules. In our experiments we take advantage of revision pairs with minimal difference between T and H, and assume that for such pairs we have only one rule to extract. We assume therefore that T-H pairs derived from Wikipedia revisions have strong similarity to the *atomic pairs* (described in Chapter 4). The actual algorithm for filtering out revision pairs with more than one phenomenon is described in Section 7.4.2.

Directionality. A Wikipedia revision, in principle, may be interpreted either as T entailing H, or as H entailing T. However, through a manual inspection of a revision sample it came out that in most of the cases the meaning of the revised sentence (T) entails the meaning of the original sentence (H). Given such observation, for the experiments reported in Sections 7.4 and 7.5 we have assumed that for all revision pairs, the revised sentence (T) entails the original sentence (H).

Context of a rule. We defined the notion of context of a rule R as a set of morpho-syntactic constraints C over the application of R in a specific T-H pair. Ideally, the set of such constraints should be the minimal set of constraints over R such that the cases of successful applications of R are maximized (e.g. the precision-recall mean is the highest). Intuitively, given an entailment rule, in absence of constraints we have the highest

recall (the rule is always applied when the Left-Hand-Side is activated in T and the Right-Hand-Side is activated in H), although we may find cases of wrong application of the rule (i.e. low precision). On the other side, as syntactic constraints are required (e.g. the subject of a verb has to be a proper name, a preposition must be followed by a prepositional phrase) the number of successful applications increase, although we may find cases where the constraints prevent the correct application (e.g. low recall).

In the absence of a data set where we could empirically estimate precision and recall of rule application, we have approximated the ideal context on the base of linguistic intuitions, defining, for different syntactic heads of the rule, the most appropriate syntactic constraints through a search algorithm over the syntactic tree produced on T and H (this is explained in detail in Section 7.4.4).

7.4 Entailment Rules Acquisition

In the next Sections, the steps for the acquisition of high precision rules from Wikipedia pairs are described in detail.

7.4.1 Step 1: Preprocessing Wikipedia dumps

As a first step, we downloaded two dumps of English Wikipedia (one dated 6.03.2009, that we will call *Wiki 09*, and one dated 12.03.2010, *Wiki 10*)[4]. We used the script *WikiExtractor.py*[5] to extract plain text from the documents, discarding any other information or annotation present in Wikipedia pages (e.g. images, tables, references and lists), but keeping the reference to the original document. Table 7.1 shows some statistics about the documents extracted from Wikipedia. For our goal, we are interested

[4] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[5] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

in the documents that are present in both *Wiki 09* and *Wiki 10* and that are not identical.

	# documents		
total Wiki 09	3 069 584		
total Wiki 10	3 038 074		
only in Wiki 09	474 117		
only in Wiki 10	505 627		
in both Wiki 09 and Wiki 10	2 563 957	identical **not ident.**	1 023 087 **1 540 870**

Table 7.1: Statistics on Wikipedia dumps.

	# pairs
set a: containment	1 547 415
set b: minor editing	**1 053 114**
set c: major editing	2 566 364

Table 7.2: Statistics on pairs similarity.

7.4.2 Step 2: Extraction of entailment pairs

For both *Wiki 09* and *Wiki 10* each document has been sentence splitted, and the sentences of the two versions have been aligned to create pairs. To measure the similarity between the sentences in each pair, we adopted the *Position Independent Word Error Rate (PER)* (Tillman *et al.* 1997 [89]), a metric based on the calculation of the number of words which differ between a pair of sentences (*diff* function in 7.1). Such measure is based on Levenshtein distance, but works at word level, and allows for re-ordering of words and sequences of words between the two texts (e.g. a translated text s and a reference translation r). It is expressed by the formula:

$$PER(s,r) = \frac{diff(s,r)+diff(r,s)}{\|r\|} \tag{7.1}$$

Setting different thresholds (T), we clustered the pairs into different sets:

- pairs composed by identical sentences, meaning that no editing was done in the more recent version of Wikipedia. Since such pairs were useless for our purposes, we discarded them. If only one word was different in the two sentences, we checked if it was a typo correction using Damerau-Levenshtein Distance (Damerau 1964 [31]). If that was the case, we discarded such pairs as well.

- pairs in which one of the sentences contains the other one, meaning that the users added some information in the new version, without modifying the old one (set a).

- pairs composed by very similar sentences, where minor editing has been carried out by the users ($PER < 0.2$) (set b). We filtered out pairs where differences were correction of misspelling and typos, and two words sentences.

- pairs composed by similar sentences, where major editing has made ($0.2 < PER < 0.6$), but still describe the same event (set c).

- pairs in which the similarity between sentences is low ($PER > 0.6$), so we discarded such pairs.

Table 7.2 shows some statistics about the extracted pairs. For our goal of extracting entailment rules, we will consider only the pairs contained in *set b*, that we consider as the atomic pairs described in Chapter 4. For each pair we intuitively set the sentence extracted from *Wiki 10* as the Text, since we assume it to have more (and more precise) information with

respect to the sentence extracted from *Wiki 09*, that we set as the Hypothesis (see Examples 7.2 and 7.3).

(7.2) T: *The Oxford Companion to Philosophy says "there is no single defining position that all anarchist hold [...]"*

H: *According to the Oxford Companion to Philosophy "there is no single defining position that all anarchist hold [...]"*

(7.3) T: *Bicycles are used by all socio-economic groups because of their convenience [...].*

H: *Bicycles are used by all socio-economic groups due to their convenience [...].*

7.4.3 Step 3: Extraction of entailment rules

All the pairs in *set b* (i.e. atomic pairs) are collected in a data set, and processed with Stanford parser (Klein and Manning 2003 [25]). Chunks have been extracted from each pair using the script *chunklink.pl*.[6] Then, we implemented algorithm 7.4.1 (and the subprocedure represented in algorithm 7.4.2), and we run them on the data sets to extract the entailment rules. The assumption is that the difference between T and H (the editing made by the user on a specific structure), can be extracted and used as entailment rule.

Algorithm 4.1-2 compares the chunks of T and H to extract the ones that differ in T and H. In details, it iteratively compares the chunks of T (*chunkT*) and H (*chunkH*), and if equal chunks are found, the algorithm checks if previous chunks are equal as well. If this is the case, these chunks are matched, and the procedure goes on. Otherwise, the algorithm searches for the unmatched chunk in H that is equal to the current chunk in T, and whose previous chunks are equal. If no matches are found, the

[6] `http://ilk.uvt.nl/team/sabine/chunklink/README.html`

current chunk from T is saved into an array (*DIF[k].chunkT*). Adjacent unmatched chunks from T are grouped together as one element of the array (*consecutive_chunkT*). Once the algorithm has iterated over each chunk from T, those chunks from H that are not matched with chunks from T, are saved into another array (*DIF[k].chunkH*) with the same id, since they were found in the same position. Adjacent unmatched chunks from H are grouped together as one element of the array (*consecutive_chunkH*).

Algorithm 7.4.1: RULES EXTRACTOR($file\_pairs, output\_file$)

```
main
 while (notEOF(file_pairs))
       do { lineT ← extract_line(file_pairs)
            lineH ← extract_line(file_pairs)
            T ← extract_chunk(lineT)
            H ← extract_chunk(lineH)
            i ← 0
            j ← 0
            k ← 0
            m ← 0

            while (i <= length(T))or(j <= length(H))
                  do { chunkT ← T[i]
                       chunkH ← T[j]
                       if (chunkT <> chunkH)
                          then { DIF[k].ID ← m
                                 DIF[k].chunkT ← chunkT
                                 DIF[k].chunkH ← chunkH
                                 k ← k + 1
                       m ← m + 1

            CREATE_RULE(output_file, DIF)
```

Algorithm 7.4.2: PROCEDURE CREATE_RULE($output\_file, DIF$)

```
procedure CREATE_RULE(output_file, DIF)
 k ← 0
 i ← 0
 while (k <= length(DIF))
       do { consecutive_chunkT ← DIF[k].chunkT
            consecutive_chunkH ← DIF[k].chunkH

            while ((DIF[k].ID + 1) = DIF[k + 1].ID)
                  do { consecutive_chunkT ← concatenate(con−
                       secutive_chunkT, DIF[k + 1].chunkT)
                       consecutive_chunkh ← concatenate(con−
                       secutive_chunkH, DIF[k + 1].chunkH)
                       k ← k + 1
            rule[i].ID ← i
            rule[i].T ← consecutive_chunkT
            rule[i].H ← consecutive_chunkH

            if found_because(rule[i].T)
             or found_because(rule[i].H)
              then {return (rule[i].ID, rule[i].T, rule[i].H)

            if found_before(rule[i].T)
             or found_before(rule[i].H)
              then {return (rule[i].ID, rule[i].T, rule[i].H)
```

Rules are therefore created setting an element from the first array (i.e. the unmatched chunks from T) as Left-Hand-Side (LHS) of the rule, and an element of the second array (i.e. the unmatched chunks from H with the same id) as the Right-Hand-Side (RHS) of the rule. The *found_because* and *found_before* functions will be explained in Section 7.5. As mentioned above, two consecutive chunks that are different in T and H are considered

to be part of the same rule (i.e. only one rule is generated for that pair). For instance, from the pair shown in Example 7.3, the rule:

Entailment rule:	**causative_1**
Pattern:	*because of* ⇒ *due to*

is extracted. On the contrary, two non consecutive chunks generate two different entailment rules.

7.4.4 Step 4: Rules expansion with minimal context

As introduced before, our work aims at providing high precision entailment rules, i.e. that they should be true any time they are applied to RTE pairs. So far, the rules extracted by algorithm 7.4.1-2 are too general with respect to our goal. For this reason, we applied algorithm 7.4.3 to add the minimum context to each rule, as discussed in Section 7.3. As input, we provide both the file with the syntactic representation of the pairs (obtained with Stanford parser), and the file with the rules extracted at Step 3. For every pair, and separately for T and H, the words isolated in the corresponding rule are matched in the syntactic tree of that sentence, and the common subsumer node is detected. Different strategies are applied to expand the rule, according to linguistic criteria. In details, if the common subsumer node is *i)* a Noun Phrase (NP) node, the rule is left as it is; *ii)* a Prepositional Phrase node (PP), all the terminal nodes of the subtree below PP are extracted; *iii)* a clause introduced by a subordinating conjunction (SBAR), all the terminal nodes of the subtree below SBAR are extracted; *iv)* an adjectival node (ADJ), all the terminal nodes of the tree below the parent of the ADJ node are extracted; *v)* a Verbal Phrase node (VP), the dependency tree under the VP node is extracted. For instance, Figure 7.1 and Figure 7.2 show the application of the algorithm to Example 7.3.

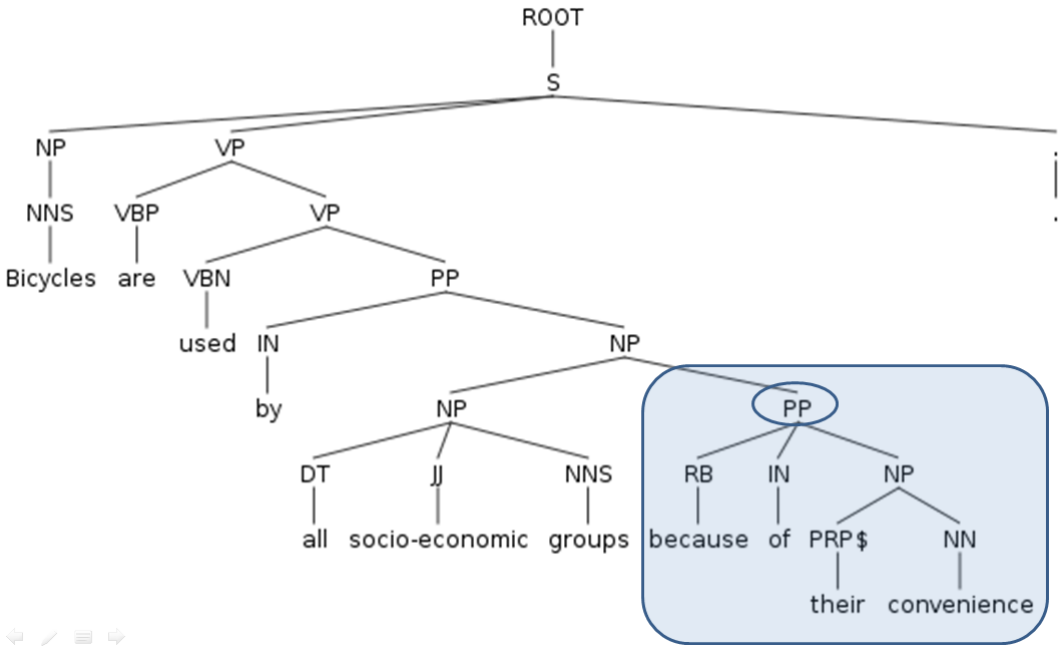


Figure 7.1: Minimal context LHS rule

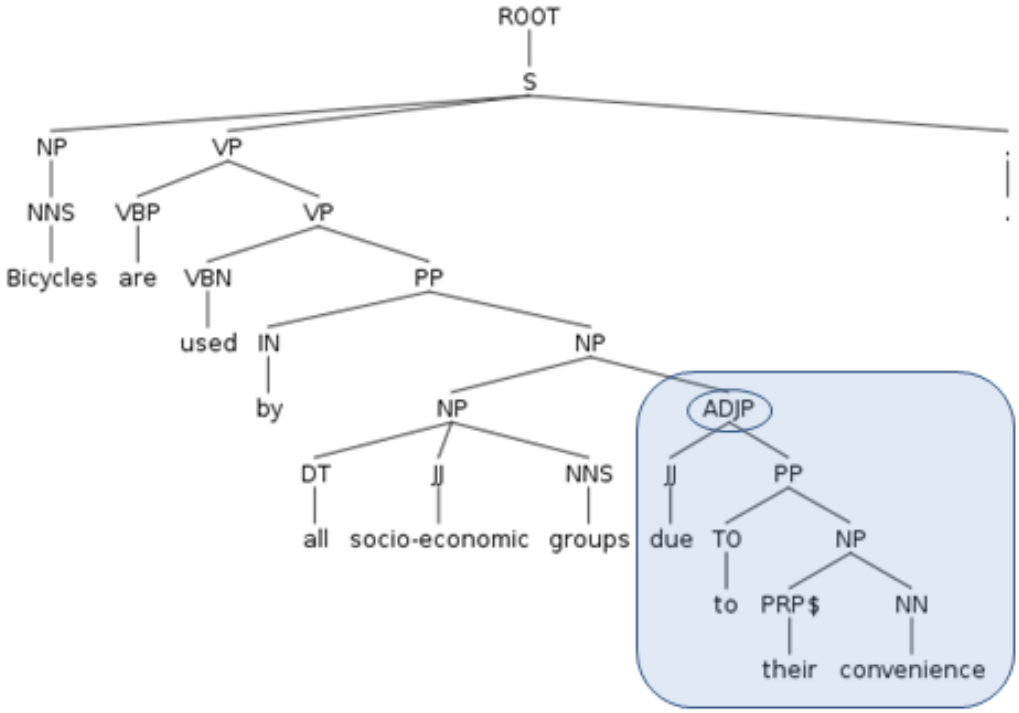


Figure 7.2: Minimal context RHS rule

Algorithm 7.4.3: EXPAND_RULE($file\_pairs, rules$)

```
main
 while (notEOF(file_pairs))
       ⎧ lineT ← read(file_pairs)
       ⎪ lineH ← read(file_pairs)
       ⎪ T_syn ← extract_syntax_tree(lineT)
       ⎪ H_syn ← extract_syntax_tree(lineH)
       ⎪ T_dep ← extract_dependency_graph(lineT)
   do  ⎨ H_dep ← extract_dependency_graph(lineH)
       ⎪
       ⎪ SEARCH_CONTEXT(rule[i].T, T_syn, T_dep)
       ⎪ SEARCH_CONTEXT(rule[i].H, H_syn, H_dep)
       ⎪
       ⎩ return (rule_expanded[i].T, rule_expanded[i].H)

procedure SEARCH_CONTEXT(rule, tree, graph)
 Common_Subs_Node ← parent_node(rule, tree)
 if (Common_Subs_Node = NP)
   then {rule_expanded ← rule

   else if (Common_Subs_Node = PP)
   then {rule_expanded ← extract_tree(PP)

   else if (Common_Subs_Node = SBAR)
   then {rule_expanded ← extract_tree(SBAR)

   else if (Common_Subs_Node = ADJ)
   then {rule_expanded ← extract_tree_parent(ADJ)

   else if (Common_Subs_Node = VP)
   then {rule_expanded ← extract_dependencies(VP)
```

The LHS of the rule *because of* is matched in the syntactic tree of T

and the prepositional phrase (PP) is identified as common subsumer node. All the terminal nodes and the PoS of the tree below PP is then extracted. The same is done for the RHS of the rule, where the common subsumer node is an adjectival phrase (ADJP).

7.5 Experiments and results

In the previous Section, we described the steps carried out to acquire high precision entailment rules from Wikipedia revision history. To show the applicability of the adopted methodology, we have performed two large-scale experiments focusing, respectively, on entailment rules for *causality* and *temporal expressions*. In particular, as case studies we chose two seeds: the conjunction *because* to derive rules for causative phrases, and the preposition *before* to derive rules for temporal expressions. For this reason, we extracted from *set b* only the pairs containing one of these two seeds (either in T or in H) and we built two separate data sets for our experiments.

While applying algorithm 4.1 we filtered again the rules acquired, collecting only those containing one of the two seeds (either in the LHS or in the RHS), using the functions *found_because* and *found_before*. This second filtering has been done because there could be pairs in which either *because* or *before* are present, but the differences in T and H do not concern those seeds. Algorithm 7.4.3 has then been applied to the selected rules to add the minimal context. The resulting rule for Example 7.3 is therefore:

Ent. rule:	**ruleid="23" docid="844" pairid="15"**		
Pattern:	*(PP*		*ADJP*
	RB 8 because) (IN 9 of)	⇒	*(JJ 8 due)(PP (TO 9 to)*
	(NP (PRP 10 their)		*(NP (PRP 10 their)*
	(NN 11 convenience)))		*(NN 11 convenience))))*

For our goals of creating entailment rules balancing high-precision with their recall (as explained in Section 7.3), when the words of the context added to the rule in Step 4 are identical we substitute them with their PoS. For Example 7.3 the rule is generalized as follows:

Ent. rule:	**ruleid="23" docid="844" pairid="15"**		
Pattern:	*(PP*		*ADJP*
	RB 8 because) (IN 9 of)	⇒	*(JJ 8 due)(PP (TO 9 to)*
	(NP (PRP)		*(NP (PRP)*
	(NN)))		*(NN))))*

The intuition underlying this generalization phase is to allow a more frequent application of the rule, while keeping some constraints on the allowed context. For instance, the application of the generalized rule from Example 2 is allowed if the subtrees below the seed words are the same (the rule can be instantiated and applied in another T-H pair as, e.g., *because of his temperament* ⇒ *due to his temperament*).

Generally, the presence of contradictions (e.g. antonyms and semantic oppositions) is really infrequent, but especially for certain cases (e.g. temporal expressions) they can have high impact (one of the most frequent rule collected for temporal expression is *before S* ⇒ *after S*). For this reason, we used WordNet to recognize antonyms, and we filtered them out during the generalization phase.

Table 7.3 shows some statistics about the resulting data sets, i.e. the numbers of acquired rules both before and after the generalization phase. All the identical rules are collapsed into a unique one, but the value of their frequency is kept in the header of that rule. Such index can then be used to estimate the correctness of the rule and, according to our intuition, the probability that the rule maintains the entailment relation.

	causality (*because*)	temporal exp. (*before*)
# rules before generalization	1671	813
# rules after generalization	**1249**	**665**
rules frequency > 2	76	40

Table 7.3: Statistics on the data sets of entailment rules.

7.5.1 Evaluation

Due to the sparseness of the phenomena under consideration (i.e. causality and temporal expressions) in RTE data sets, evaluating the acquired sets of entailment rules on such data would not provide interesting results from the point of view of the quality of the extracted rules.

For this reason, we opted for a manual analysis of a sample of 100 rules per set, including all the rules whose frequency was higher than 2 (see Table 7.3), plus a random set of rules whose frequency was equal to 1. In the analysis we carried out, we differentiate three possible values for a rule: *entailment=yes* (i.e. correctness of the rule); *entailment=no* (meaning that the entailment relation does not hold between the LHS and the RHS of the rule, often because the editing has changed the semantics of the proposition); *entailment=no-error* (i.e. the rule is wrong, either because the editing in *Wiki10* was done to correct mistakes, or because the rule is not well-formed due to mistakes produced by Algorithm 7.4.1-2). Table 7.4 shows the results of the analysis, discussed in the next Section.

	freq > 2	% yes	% no	% error
causality	76	72	15	13
temporal exp.	33	51	21	28

Table 7.4: Results of the evaluation of the sets of rules.

7.5.2 Error Analysis

In general, due to the amount of noisy data present in Wikipedia, incorrect rules can be collected. Some editing done by the users can indeed be spelling, typographical or ungrammaticality corrections, or just spam. Analysing the sample of the rules manually, we found out that noisy rules are about 10% of the total. Some spell-checker or dictionary-based filters should be applied to automatically discard some of these cases.

As introduced before, another strategy to select only reliable rules is to use the frequency with whom they can be found in the data, to estimate the confidence that a certain rule maintains the entailment. Since the procedure to create the rules privileges their precision, only a few rules appear very frequently (especially for temporal expressions, as shown in Table 7.3), and this can be due to the constraints defined for the context extraction. This fact motivates also the lower precision of the rules for temporal expressions, where 77% of the sample we analysed involved rules with frequency equal to 1. Furthermore, most of the rules we annotated as *entailment=no* are due to the fact that the editing of *Wiki10* concerned a change in the semantics of the pair, resulting into the *unknown* judgement. Examples of this kind are for instance, the rule: *before 1990* $\Rightarrow$ *1893* for temporal expressions, or *when x produced* $\Rightarrow$ *because x produced*. Defining and experimenting further strategies to empirically estimate precision and recall of rules application are needed, and are part of future work. Indeed, several rules that appear only once represent good rules, and should not be discarder a priori.

Finally, the idea of using only very similar pairs to extract entailment rules is based on the assumption that such rules should concern one phenomenon at a time (as suggested in Bentivogli *et al.* 2010 [11]). Despite the strategies adopted to avoid to have more than one phenomenon per

rule, in about 10% of the cases two phenomena (e.g lexical and syntactic) are collapsed on consecutive tokens, so it was not possible to separate them automatically (e.g. *because of the divorce settlement cost* ⇒ *due to the cost of his divorces settlement*, where the causative and the argument realization rules should be separated).

7.6 Conclusion and future work

In Chapter 6 we carried out a pilot study on RTE pairs to isolate the phenomena relevant to the entailment relation, with the goal of creating atomic T-H pairs to allow TE systems training and evaluation on specific inference types. To create atomic pairs, an entailment rule is individuated for a certain phenomenon, and it is instantiated using the portion of T which expresses that phenomenon as the LHS of the rule, and information from H on the same phenomenon as the RHS of the rule.

While that pilot study has been manually performed to become aware of the difficulties and the problems of the task, in this Chapter we have presented a methodology for the automatic acquisition of entailment rules from Wikipedia revision pairs. The main benefits are the following: *i)* large-scale acquisition, given the size of Wikipedia revisions (continuously increasing); *ii)* new coverage, because Wikipedia revisions contain linguistic phenomena (e.g. causality, temporal expressions), which are not covered by existing resources: as a consequence we can significantly extend the coverage of current TE systems; *iii)* quality, we have introduced a novel notion of context of the rule, based on the minimal set of syntactic features that maximize the successful application of the rule, and have implemented it as a search over the syntactic representation of revision pairs.

The results obtained on two experimental acquisitions, respectively on causality (using the seed *because*) and temporal expressions (using the seed

before) show both a very high quality and coverage of the extracted rules. The obtained resource includes, respectively, 1249 and 665 rules, which cover entailment and paraphrasing aspects not represented in other similar resources. Since the methodology does not require human intervention, the resource can be easily extended applying the algorithms to collect rules for other phenomena relevant to inference; furthermore, it can be periodically updated, as Wikipedia revisions change continuously. [7]

Since in our component-based framework the entailment rules define the allowed transformation (atomic edits) for a certain phenomenon relevant to inference, having a strategy to automatically collect them is of great value. The results we obtained in our study encourage us to further improve the approach, considering a number of directions. First, we plan to improve the capacity to filter out revision pairs that contain more than one phenomenon (step 2 of the procedure in Section 7.4): this might be obtained again considering the syntactic structure of the sentence. Second, we plan to couple the use of frequency filters with the use of typical contradiction patterns (e.g. use of negation, antonyms derived from WordNet) in order to detect revision pairs with contradictory information.

Finally, we are planning more extended evaluations, which include the integration of the extracted rules into existing TE systems. However, this evaluation has to be carefully designed, as the ablation tests carried on at RTE show. In particular, as RTE tasks are moving toward real applications (e.g. summarization, as described in Chapter 3) we think that knowledge reflecting real textual variations produced by humans (as opposed to knowledge derived from linguistic resources) may introduce interesting and novel hints.

[7] The resources we created, as well as new extensions, are freely available at `http://hlt.fbk.eu/en/technology`

Chapter 8

Component-based Evaluation of Textual Entailment Systems

This Chapter presents a methodology for Textual Entailment systems evaluation, that takes advantage of the decomposition of Text-Hypothesis pairs into atomic pairs (as described in Chapter 6) and propose to run systems over such data sets. As a result, a number of quantitative and qualitative indicators about strength and weaknesses of TE systems are highlighted. As a pilot study, we evaluate and compare three TE-systems, namely EDITS, VENSES ans BLUE, basing on this methodology.

8.1 Introduction

The intuition underlying the component-based framework for TE we propose in this Thesis, is that the more a system is able to correctly solve the linguistic phenomena relevant to the entailment relation separately, the more the system should be able to correctly judge more complex pairs, in which different phenomena are present and interact in a complex way. As discussed in Chapter 4, such intuition is motivated by the notion of meaning compositionality, according to which the meaning of a complex expression is determined by its structure and by the meaning of its con-

stituents (Frege 1992 [37]). In a parallel way, we assumed that it is possible to recognize the entailment relation of a T-H pair only if all the phenomena contributing to such a relation (i.e. the atomic arguments) are resolved.

Remaining faithful to these assumptions, we reasoned about the advantages of exploiting the procedure to decompose complex pairs into atomic arguments (presented in Chapter 4), to define an evaluation framework that could offer an insight into the kinds of sub-problems a given system can reliably solve. The metric that is currently used to evaluate TE system performances, i.e. accuracy, turns out to be opaque, and inadequate to assess systems capabilities in details. Experiments like the ablation tests attempted in the last RTE-5 and RTE-6 campaigns on lexical and lexical-syntactic resources go in this direction, although the degree of comprehension is still far from being optimal, as discussed in Chapter 3.

Basing on our study on the atomic arguments that are relevant for inference (described in Chapter 4), in this Chapter we propose a *component-based evaluation*, that aims at providing a number of quantitative and qualitative indicators about a TE system. Evaluation is carried out both on the original T-H pairs and on the atomic pairs originated from it (Section 8.2). This strategy allows to analyse the correlations among the capability of a system to address single linguistic phenomena in a pair and the ability to correctly judge the pair itself. Despite the strong intuition about such correlation (i.e. the more the phenomena for which a system is trained, the better the final judgement), no empirical evidence support it yet.

For this reason we carried out a pilot study, testing the component-based method on a sample of 60 pairs - extracted from the resource described in Chapter 6 - each decomposed in the corresponding atomic pairs, and using three systems that obtained similar performances in RTE-5 (Section 8.3). The main features and differences of these systems come to light when evaluated using qualitative criteria. Furthermore, we compare such

systems with two different baseline systems, the first one performing Word Overlap, while the second one is an ideal system that knows *a priori* the probability of a linguistic phenomenon to be associated with a certain entailment judgement.

Moreover, we investigate the correlations defined above on different subsets of the evaluation data set (i.e. positive vs negative pairs) and we try to induce regular patterns of evaluation, to understand if some phenomena are more relevant for a certain judgement rather than for another (Section 8.4). In particular, we carried out an analysis on contradiction judgements, highlighting *i)* the variety of linguistic phenomena that are relevant for such judgement, and *ii)* how polarity among Text and Hypothesis affects the entailment/contradiction judgements (e.g. whether specific combinations of phenomena are more frequent than others).

8.2 Component-based evaluation

Aim of the component-based evaluation described in this Section is to provide quantitative and qualitative indicators about the behaviours of actual TE systems. In particular, in the component-based system proposed in this Thesis, such methodology allows to independently evaluate the TE-components, and to assess the impact of their performances on the final result.

8.2.1 General Method

As introduced before, the evaluation methodology we propose assumes Fregean meaning compositionality principle. According to such assumption, we expect that the higher the accuracy of a system on the atomic pairs (as defined in Chapter 4) and the compositional strategy, the better its performances on the original RTE pairs. Moreover, the precision a

system gains on single phenomena should be maintained over the general data set, thanks to suitable mechanisms of meaning combination.

Given a data set composed of original RTE pairs [T-H], a data set composed of all the atomic pairs derived from it [T-H]$_{atomic}$, and a TE system S, the evaluation methodology we propose consists of the following steps:

1. Run S both on [T-H] and on [T-H]$_{atomic}$, to obtain the accuracies of S both on the RTE original and on the atomic pairs;

2. Extract data concerning the behaviour of S on each phenomenon or on classes of phenomena, and calculate separate accuracies. This way it is possible to evaluate how much a system is able to correctly deal with single or with classes of phenomena;

3. Calculate the correlation between the ability of the system to correctly judge the atomic pairs of [T-H]$_{atomic}$ with respect to the ability to correctly judge the original ones in [T-H]. Such correlation is expressed through a *Component Correlation Index (CCI)*, as defined in Section 8.2.2;

4. In order to check if the same *CCI* is maintained over both entailment and contradiction pairs (i.e. to verify if the system has peculiar strategies to correctly assign both judgements, and if the high similarity of atomic pairs does not bias its behaviour), we calculate a *Component Deviation Index (CDI)* as the difference between the *CCIs* on entailment and on contradiction pairs, as explained in more details in Section 8.2.3.

8.2.2 Component Correlation Index (CCI)

We assume that the accuracy obtained on [T-H]$_{atomic}$ should positively correlate with the accuracy obtained on [T-H]. We define a *Component Correlation Index* as the ratio between the accuracy of the system on the original RTE data set and the accuracy obtained on the atomic data set, as follows:

$$CCI = \frac{acc[T-H]}{acc[T-H]_{atomic}} \tag{8.1}$$

We expect the component correlation index of an optimal ideal system (or the human goldstandard) to be equal to 1, i.e. 100% accuracy on the atomic data set should correspond to 100% accuracy on the original RTE data set. For this reason, we consider $CCI = 1$ as the ideal correlation, and we calculate the difference between such ideal CCI and the correlation obtained for a system S. Given such expectations, CCI_S can assume three different configurations with respect to the upperbound (i.e. the ideal correlation):

- $CCI_S \cong 1$ (ideal correlation): when CCI_S approaches to 1, the system shows high correlation with the ideal behaviour assumed by the compositionality principle. As a consequence, we can predict that improving single modules will correspondingly affect the global performance.

- $CCI_S < 1$ (missing correlation): the system is not able to exploit the ability in solving single phenomena to correctly judge the original RTE pairs. This may be due to the fact that the system does not adopt suitable combination mechanisms and loses the potentiality shown by its performances on atomic pairs.

- $CCI_S > 1$ (over correlation): the system does not exploit the ability to solve atomic arguments to solve the whole pairs, and has different mechanisms to evaluate the entailment. Probably, such a system is not intended to be modularized.

Beside this "global" component correlation index calculated on the complete RTE data and on all the atomic pairs created from it, the *CCI* can also be calculated *i)* on categories of phenomena, to verify which phenomena a system is more able to solve both when isolated and when interacting with other phenomena, e.g.:

$$CCI_{lex} = \frac{acc[T-H]_{lex}}{acc[T-H]_{atomic-lex}} \tag{8.2}$$

including in [T-H]$_{lex}$ all the pairs in which at least one lexical phenomenon is present and contribute to the entailment/contradiction judgements, and in [T-H]$_{atomic-lex}$ all the atomic pairs in which a lexical phenomenon is isolated; or *ii)* on kind of judgement (*entailment*, *contradiction*, *unknown*), allowing deeper qualitative analysis of the performances of a system.

8.2.3 Component Deviation Index (CDI)

We explained that a low *CCI* (i.e. < 1) of a system reflects the inability to correctly exploit the potentially promising results obtained on atomic pairs to correctly judge RTE pairs. Actually, it could also be the case that the system does not perform a correct combination because even the results got on the atomic pairs were accidental (e.g. a word overlap system performs well on atomic pairs because of the high similarity between T and H, and not because it has linguistic competences).

We detect such cases by decomposing the evaluation data sets, separating positive (i.e. *entailment*) from negative (i.e. *contradiction*, *unknown*)

examples both in [T-H] and in $[\text{T-H}]_{atomic}$, and independently run the system on the new data sets. Then, we have finer grained evaluation patterns through which we can analyse the system behaviour.

In the ideal case, we expect to have good correlation between the accuracy obtained on the atomic pairs and the accuracy obtained on the original ones ($0 < CCI_{pos} \leq 1$ and $0 < CCI_{neg} \leq 1$). On the contrary, we expect that systems either without a clear composition strategy or without strong components on specific linguistic phenomena (e.g. a word overlap system), would show a significant difference of correlation on the different data sets. More specifically, situations of *inverse correlation* on the entailment and contradiction pairs (e.g. over correlation on contradiction pairs and missing correlation on entailment pairs) may reveal that the system itself is affected by the nature of the data set (i.e. its behaviour is biased by the high similarity of $[\text{T-H}]_{atomic}$), and weaknesses in the ability of solving phenomena that more frequently contribute to the assignment of a contradiction (or an entailment) judgement come to light.

We formalize such intuition defining a *Component Deviation Index (CDI)* as the difference between the correlation indexes, respectively, on entailment and contradiction/unknown pairs, as follows:

$$|CDI| = CCI_{pos} - CCI_{neg} \tag{8.3}$$

For instance, a high Component Deviation Index due to a missing correlation on positive entailment pairs and an over correlation for negative pairs, is interpreted as an evidence that the system has low accuracy on $[\text{T-H}]_{atomic}$ - T and H are very similar and the system has no strategies to understand that the phenomenon that is present must be judged as contradictory -, and a higher accuracy on [T-H], probably due to chance. In the ideal case $CDI_S \cong 0$, since we assumed the ideal *CCIs* on both positive

and negative examples to be as close as possible to 1 (see Section 8.2.2).

8.3 Experiments and discussion

This Section describes the experimental setup of our pilot study, carried out using three systems that took part in RTE-5, i.e. EDITS, VENSES, and BLUE. We show the results obtained and the analysis performed basing on the proposed component-based evaluation. Their respective *CCIs* and *CDIs* are compared with two baselines: a word overlap system, and a system biased by the knowledge of the probability that a linguistic phenomenon contributes to the assignment of a certain entailment judgement.

8.3.1 Data set

The evaluation method has been tested on a data set composed of 60 pairs from RTE-5 test set ($[\text{T-H}]_{RTE5-sample}$, composed of 30 *entailment*, and 30 *contradiction* randomly extracted examples), and a data set composed of all the atomic pairs derived by the first one (we used the resource described in Chapters 4 and 6). This second data set $[\text{T-H}]_{RTE5-atomic}$ is composed of 167 pairs (135 *entailment*, 32 *contradiction* examples, considering 35 different linguistic phenomena - listed in Chapter 4). In this pilot study we decided to limit our analysis to entailment and contradiction pairs since, as observed in Chapter 6, in most of the unknown pairs no linguistic phenomena relating T to H can be detected.

8.3.2 TE systems

EDITS The EDITS system (Edit Distance Textual Entailment Suite) (Kouylekov and Negri 2010 [49]) has been described in details in Chapter 5. For our experiments we applied the model that produced EDITS best

run at RTE-5 (2 way, accuracy on test set: 60.2%) (Mehdad *et al.* 2009 [64]). The main features are: Tree Edit Distance algorithm on the parsed trees of T and H, Wikipedia lexical entailment rules, and PSO optimized operation costs (Mehdad 2009 [61]).

VENSES The second system used in our experiments is VENSES[1] (Delmonte *et al.* 2009 [33]), that obtained performances similar to EDITS and BLUE at RTE-5 (2 way, accuracy on test set: 61.5%). It applies a linguistically-based approach for semantic inference, and is composed of two main components: *i)* a grammatically-driven subsystem validates the well-formedness of the predicate-argument structure and works on the output of a deep parser producing augmented head-dependency structures; and *ii)* a subsystem detects allowed logical and lexical inferences basing on different kind of structural transformations intended to produce a semantically valid meaning correspondence. Also in this case, we applied the best configuration of the system used in RTE-5.

BLUE The third system experimented is BLUE (Boeing Language Understanding Engine) (Clark and Harrison 2009 [23]), that is based on a "logical" approach to RTE. It first creates a logic-based representation of a text T and then performs simple inference (using WordNet and the DIRT inference rule database) to try and infer an hypothesis H. The overall system can be viewed as composed by three main elements: parsing, WordNet, and DIRT, built on top of a simple baseline of bag-of-words comparison. BLUE's best score on at RTE-5 is 61.5% (2 way) and 54.7% (3 way).

Baseline system 1: Word Overlap algorithm The first baseline applies a Word Overlap (WO) algorithm on tokenized text. The threshold to sepa-

[1] http://project.cgm.unive.it/venses_en.html

rate positive from negative pairs is learnt on RTE-5 training data set.

Baseline system 2: Linguistic biased system The second baseline is produced by a more sophisticated but biased system. It exploits the probability of linguistic phenomena to contribute more to the assignment of a certain judgement than to another. Such probabilities are learnt on the $[\text{T-H}]_{RTE5-atomic}$ goldstandard: given the list of the phenomena with their frequency in atomic positive and negative pairs (columns 1,2,3 of Table 8.1), we calculate the probability P of phenomenon i to appear in a positive (or in a negative) pair as follows:

$$P(i|[T-H]_{positive}) = \frac{\#(i|[T-H]_{RTE5-positive-atomic})}{\#(i|[T-H]_{RTE5-atomic})} \tag{8.4}$$

For instance, if the phenomenon *apposition* appears in 11 atomic positive pairs and in 6 negative pairs, it has a probability of 64.7% to appear in positive examples and 35.3% to appear in negative ones. Such knowledge is then stored in the system, and is used in the classification phase, assigning the most probable judgement associated to a certain phenomenon.

When applied to $[\text{T-H}]_{RTE5-sample}$, this system uses a simple combination strategy: if phenomena associated with different judgements are present in a pair, and one phenomenon is associated with a contradiction judgement with a probability $> 50\%$, the pair is marked as *contradiction*, otherwise it is marked as *entailment*.

8.3.3 Results

Following the methodology described in Section 8.2, at step 1 we run EDITS, VENSES and BLUE on $[\text{T-H}]_{RTE5-sample}$, and on $[\text{T-H}]_{RTE5-mono}$ (Table 8.2 reports the accuracies obtained).

phenomena	# [**T-H**] *RTE5−atomic*		**EDITS** % acc.		**VENSES** % acc.		**BLUE** % acc.	
	pos.	**neg.**	**pos.**	**neg.**	**pos.**	**neg.**	**pos.**	**neg.**
lex:identity	1	3	100	0	100	33.3	100	0
lex:format	2	-	100	-	100	-	100	-
lex:acronymy	3	-	100	-	33.3	-	100	-
lex:demonymy	1	-	100	-	100	-	100	-
lex:synonymy	11	-	90.9	-	90.9	-	100	-
lex:semantic-opp.	-	3	-	0	-	100	-	33.3
lex:hypernymy	3	-	100	-	66.6	-	66.6	-
lex:geo-knowledge	1	-	100	-	100	-	100	
TOT lexical	22	6	95.4	0	77.2	66.6	95.22	16.65
lexsynt:transp-head	2	-	100	-	50	-	50	-
lexsynt:verb-nom.	8	-	87.5	-	25	-	50	-
lexsynt:causative	1	-	100	-	100	-	100	-
lexsynt:paraphrase	3	-	100	-	66.6	-	66.6	-
TOT lex-syntactic	14	-	92.8	-	42.8	-	66.65	-
synt:negation	-	1	-	0	-	0	-	100
synt:modifier	3	1	100	0	33.3	100	100	-
synt:arg-realization	5	-	100	-	40	-	80	-
synt:apposition	11	6	100	33.3	54.5	83.3	90.9	33.3
synt:list	1	-	100	-	100	-	0	-
synt:coordination	3	-	100	-	33.3	-	66.6	-
synt:actpass-altern.	4	2	100	0	25	50	100	-
TOT syntactic	28	9	96.4	22.2	42.8	77.7	72.9	66.6
disc:coreference	20	-	95	-	50	-	90	-
disc:apposition	3	-	100	-	0	-	100	-
disc:anaphora-zero	5	-	80	-	20	-	100	-
disc:ellipsis	4	-	100	-	25	-	100	-
disc:statements	1	-	100	-	0	-	0	-
TOT discourse	33	-	93.9	-	36.3	-	78	-

reas:apposition	2	1	100	0	50	100	50	100
reas:modifier	3	-	66.6	-	100	-	66.6	-
reas:genitive	1	-	100	-	100	-	100	-
reas:relative-clause	1	-	100	-	0	-	100	-
reas:elliptic-expr.	1	-	100	-	0	-	100	-
reas:meronymy	1	1	100	0	100	0	100	0
reas:metonymy	3	-	100	-	33.3	-	100	-
reas:representat.	1	-	100	-	0	-	100	-
reas:quantity	-	5	-	0	-	80	-	40
reas:spatial	1	-	100	-	0	-	100	-
reas:gen-inference	24	10	87.5	50	37.5	90	75	50
TOT reasoning	38	17	89.4	35.2	42.1	82.3	89.16	47.5
TOT (all phenom)	135	32	93.3	25	45.9	81.2	80.38	43.5

Table 8.1: Systems' accuracy on phenomena

At step 2, we calculate the accuracy of EDITS, VENSES and BLUE on each single linguistic phenomenon, and on categories of phenomena. Table 8.1 shows the distribution of the phenomena in the data set, reflected in the number of positive and negative atomic pairs created for each phenomenon. As can be seen, some phenomena appear more frequently than others (e.g. *coreference*, *general inference*). Furthermore, some linguistic phenomena allow only the creation of positive or negative examples, while others can contribute to the assignment of both judgements. Due to the small data sets we used, some phenomena appear rarely; the accuracy on them cannot be considered completely reliable.

Nevertheless, from these data the main features of the systems can be identified. For instance, EDITS obtains the highest accuracy on positive atomic pairs, while it seems it has no peculiar strategies to deal with phenomena causing contradiction (e.g. *semantic opposition*, and *quantity mismatching*). Also BLUE shows the same tendency in better solving

entailment pairs with respect to contradiction pairs, even if the gap in performances is narrower. On the contrary, VENSES shows an opposite behaviour, obtaining the best results on the negative cases.

At step 3 of the proposed evaluation methodology, we calculate the correlation index between the ability of the system to correctly judge the atomic pairs of $[\text{T-H}]_{RTE5-atomic}$ with respect to the ability to correctly judge the original ones in $[\text{T-H}]_{RTE5-sample}$.

	acc. % $RTE5-sample$	acc. % $RTE5-atomic$	*CCI*	Δ
EDITS	58.3	80.8	0.72	0.28
VENSES	60	52.6	1.15	0.15
BLUE	55.9	70.2	0.78	0.22
Word Overlap	38.3	77.24	0.49	0.51
ling baseline	68.3	86.8	0.79	0.21

Table 8.2: Evaluation on RTE pairs and on atomic pairs

Table 8.2 compares EDITS, VENSES and BLUE *CCI* with the two baseline systems described before. As can be noticed, even if EDITS *CCI* outperforms the WO system, it shows a similar behaviour (high accuracy on atomic pairs, and much lower on the RTE sample). According to our definition, their *CCIs* ($0 < CCI < 1$) show a good ability of the systems to deal with linguistic phenomena when isolated, but a scarce ability in combining them to assign the final judgement. EDITS *CCI* is not far from the *CCI* of the linguistic biased baseline system, even if we were expecting a higher *CCI* for the latter system. The reason is that beside the linguistic phenomena that allow only the creation of negative atomic pairs, all the phenomena that allow both judgements have a higher probability to contribute to the creation of positive atomic pairs.

		categories of linguistic phenomena				
	RTE5 data	lex.	lex-synt.	synt.	disc.	reas.
EDITS	*sample*	47.8	64.3	51.7	75	62.5
	atomic	75	92.8	78.3	93.9	72.7
	CCI	**0.63**	**0.69**	**0.66**	**0.79**	**0. 85**
VENSES	*sample*	47.2	42.8	62	46.4	67.5
	atomic	75	42.8	51.3	33	54.5
	CCI	**0.62**	**1**	**1.2**	**1.4**	**1.23**
BLUE	*sample*	50	50	51.7	48.1	61
	atomic	78.5	50	71	87.5	69
	CCI	**0.63**	**1**	**0.72**	**0.54**	**0.88**
WO	*sample*	36.3	57.1	34.4	50	35
baseline	*atomic*	78.5	71.4	72.9	96.9	69
	CCI	**0.46**	**0.79**	**0.47**	**0.51**	**0.5**
ling-	*sample*	82.6	92.8	58.6	82.1	70
biased	*atomic*	96.4	100	75.6	96.9	80
baseline	**CCI**	**0.85**	**0.92**	**0.77**	**0.84**	**0.87**

Table 8.3: Evaluation on categories of phenomena

Comparing the *CCI* of the five analysed systems with the ideal correlation ($CCI_S \cong 1$, see Section 8.2.2), VENSES is the closest one ($\Delta = 0.15$), even if it shows a light over correlation (probably due to the nature of the data set). The second closest one is the linguistic biased system ($\Delta = 0.21$), showing that the knowledge of the most probable judgement assigned to a certain phenomenon can be a useful information.

Table 8.3 reports an evaluation of the five systems on categories of linguistic phenomena.

To check if the same *CCI* is maintained over both entailment and contradiction pairs, we calculate a *Deviation Index* as the difference between the *CCIs* on entailment and on contradiction pairs (step 4 of our methodology). As described in Section 8.2, we created four data sets dividing

both [T-H]$_{RTE5-sample}$ and [T-H]$_{RTE5-atomic}$ into positive (i.e. *entailment*) and negative (i.e. *contradiction*) examples. We run EDITS, VENSES and BLUE on the data sets and we calculate the *CCI* on positive and on negative examples separately. If we obtained missing correlation between the accuracy on the atomic pairs and the accuracy on RTE original ones, it would mean that the potentiality that the systems show on atomic pairs is not exploited to correctly judge more complex pairs, therefore compositional mechanisms should be improved.

		% acc. $_{\substack{RTE5 \\ sample}}$	% acc. $_{\substack{RTE5 \\ atomic}}$	*CCI*	*CDI*
EDITS	E	83.3	94.7	0.88	0.5
	C	33.3	24	1.38	
VENSES	E	50	47.01	1.08	0.16
	C	70	75.7	0.92	
BLUE	E	66.33	82.5	0.80	0
	C	46.66	57.9	0.80	
WO	E	50	88	0.56	0.24
baseline	C	26.6	33	0.80	
ling-biased	E	96.6	98.5	0.98	0.03
baseline	C	40	39.4	1.01	

Table 8.4: Evaluation on entailment and contradiction pairs

Table 8.4 shows that the *CDIs* of BLUE and of VENSES are close to the ideal case ($CDI_S \cong 0$), indicating a good capacity to correctly differentiate entailment from contradiction cases. EDITS results demonstrate that the shallow approach implemented by the system has no strategies to correctly judge negative examples (similarly to the WO system), therefore should be mainly improved with this respect.

8.4 Contradiction-focused analysis

The analysis we carried out in the previous Sections has shown that systems turn out to have more difficulties in assigning the correct judgement to contradiction pairs with respect to entailment pairs. This is supported by previous studies (e.g. de Marneffe *et al.* 2008 [59], Harabagiu *et al.* 2006 [43]), that claim that detecting contradiction appears to be a harder task that detecting entailment, since it is not sufficient to highlight mismatching information between sentences, but deeper comprehension is required. In RTE task, contradiction is said to occur when two sentences are extremely unlikely to be true simultaneously; furthermore, they must involve the same event. For applications in information analysis, it can be very important to detect incompatibility and discrepancies in the description of the same event, and the contradiction judgement in the TE task aims at covering this aspect. Unlike in traditional semantic analysis, in TE it is not enough to detect the polarity of a sentence, but rather it is necessary to analyse the dependencies between two sentences (T-H pair) in order to establish whether a contradiction holds between them.

Moreover, as already pointed out in (Wang *et al.* 2009 [96]), the similarity between T's and H's in pairs marked as entailment and contradiction is much higher with respect to the similarity between T's and H's in pairs marked as unknown. To support this intuition, (Bentivogli *et al.* 2009 [14]) provide some data on the lexical overlap between T's and H's in the last RTE Challenges. For instance, in RTE-4 the lexical overlap is 68.95% in entailment pairs, 67.97% in contradiction pairs and only 57.36% in the unknown pairs. Similarly, in RTE-5 the lexical overlap between T's and H's is 77.14% in entailment pairs, 78.93% in contradiction pairs and only 62.28% in the unknown pairs.

Analysing RTE data of the previous challenges, we noticed that the

tendency towards longer and more complex sentences in the data sets in order to reproduce more realistic scenarios, is also reflected in more complex structures determining contradictions. For instance, contradictions arising from overt negation as in Example 8.5 (pair 1663, RTE-1 test set):

(8.5) T: *All residential areas in South Africa are segregated by race and no black neighbourhoods have been established in Port Nolloth.*

H: *Black neighbourhoods are located in Port Nolloth.*

are infrequent in the data sets of more recent RTE challenges. For instance, in RTE-5 test set, only in 4 out of 90 contradiction pairs an overt negation is responsible for the contradiction judgement. In agreement with (De Marneffe *et al.* 2008 [59]), we also remarked that most of the contradictions involve numeric mismatch, wrong apposition, entity mismatch and, above all, deeper inferences depending on background and world knowledge, as in Example 8.6 (pair 567, RTE-5 test set):

(8.6) T: *"[...] we've done a series of tests on Senator Kennedy to determine the cause of his seizure. He has had no further seizures, remains in good overall condition, and is up and walking around the hospital".*

H: *Ted Kennedy is dead.*

These considerations do not mean that overt negations do not appear in the RTE pairs. On the contrary, they are often present in T-H pairs, but most of the times their presence is irrelevant in the assignment of the correct entailment judgement to the pair. For instance, the scope of the negation can be a phrase or a sentence with additional information with respect to the relevant parts of T and H that allow to correctly judge the pair. This fact could be misleading for systems that do not correctly exploit

syntactic information, as the experiments using Linear Distance we carried out for our participation at RTE-4 (Cabrio *et al.* 2008 [18]).

In the analysis of the distribution of the linguistic phenomena we carried out both in Chapter 6 and in the previous Sections, we noticed that due to their nature some phenomena are strongly related to a certain judgement (e.g. semantic opposition), while other appear both in positive and in negative pairs. Learning such correlations on larger data sets could be an interesting feature to be exploited by TE systems in the assignment of a certain judgement if a specific phenomenon is detected in the pair (see the linguistic-biased baseline system experimented in Section 8.3).

Table 8.5 reports the cooccurrences of the linguistic phenomena relevant to inference in the pairs marked as *contradiction*. On the first horizontal row all the phenomena that at least in one pair determine contradiction are listed, while in the first column there are all the phenomena co-occurring with them in the pairs. The idea underlying this table is to understand if it is possible to identify recurrent patterns of co-occurrences between phenomena in contradiction pairs. As can be noticed, almost all phenomena occur together with expressions requiring deeper inference (*reas:general_inference*), but this is due to the fact that this category is the most frequent one. Beside this, it seems that no specific patterns can be highlighted, but it could be worth extending this analysis increasing the number of pairs of the sample.

8.5 Conclusion

In this Chapter we have described a component-based methodology for the evaluation of TE systems, based on the analysis of the system behaviour on atomic pairs with respect to the behaviour on corresponding original pairs. Through the definition of two indicators, a Component Correlation Index

	lex:identity	lex:sem_opposition	synt:negation	synt:modifier	synt:apposition	synt:actpass_altern	reas:meronymy	reas:quantity	reas:gen_inference
lex:identity								1	1
lex:format								1	
lex:acronymy					1				
lex:synonymy	1					1		1	1
lex:hypernymy								1	
lexsynt:vrb-nom	1	1						1	
lexsynt:caus.								1	
synt:modifier									1
synt:arg-realiz.		1							1
synt:apposition		2							3
synt:coord.								1	
synt:actpass	1	1							
disc:coref.	3				1				4
disc:apposition									
disc:anaph-0							1	1	
disc:ellipsis	1	1							2
disc:statements									1
reas:genitive			1						
reas:meronymy								1	
reas:gen-infer.	1			1	3		1	2	1

Table 8.5: Cooccurrencies of phenomena in contradiction pairs

and a Component Deviation Index, we infer evaluation patterns which indicate strengths and weaknesses of the system. With respect to accuracy, the traditional way to evaluate system performances in RTE Challenges, the component-based evaluation methodology allows a more detailed assessment of system capabilities, and allow TE system developers to independently evaluate modules and algorithms implemented to cope with specific inference types. As a pilot study, we have compared three systems that took part in RTE-5. We discovered that, although the three systems have similar accuracies on RTE-5 data sets, they show significant differences in their respective abilities to manage different linguistic phenomena and to properly combine them.

Chapter 9

Conclusion

This Thesis presents and discusses the more relevant results of our research on Component-Based Textual Entailment. The framework described aims at providing a model to decompose the complexity of the Textual Entailment problem, assuming Fregean meaning compositionality principle. Several dimensions of this framework have been investigated and experimented.

First of all, we defined the main features of the proposed TE architecture composed by TE-components, each of which able to address a TE task on a specific phenomenon relevant to inference in isolation. We took advantage of the conceptual and formal tools available from an extended model of Natural Logic, to define clear strategies for their combination. In a transformation-based framework, each component performs atomic edits to process a certain linguistic phenomenon, and assigns an entailment relation as the output of this operation. NL mechanisms of semantic relations composition are then applied to join the output of each single component, in order to obtain a global entailment judgement for a pair. With respect to the model described in (Mac Cartney and Manning 2009 [56]) in which a lot of effort is made to establish the proper projectivity signatures for a broad range of quantifiers, implicative and factives, and other semantic re-

lation, our work is less fine-grained, since it relies on the expressivity of the entailment rules to model a certain linguistic phenomenon. On the other hand, as far as a linguistic phenomenon can be expressed through entailment rules it can be modelled in our framework, guaranteeing a broader coverage on RTE problems.

As a second task, we implemented a set of TE-components basing on EDITS system's modular architecture. Even if such package was not developed within this Thesis work, we provided valuable contributions to its improvement, and we adapted its architecture to account for the properties of the components previously described. Each component has been carefully shaped to reward its precision, so that it focuses only on the phenomenon it is built to deal with to avoid overlapping, and to prevent that in the entailment composition phase errors made in the initial steps would propagate to the system's final output. Part of this implementation work has been carried out during a six-month research internship at Xerox Research Center Europe, where we developed also a plug-in to adapt the dependency representation of the sentences provided by XIP (Xerox Incremental Parser) to EDITS input format, in particular to take advantage of XIP internal coreference resolver module. Such architecture has been evaluated in our participations to RTE campaigns (in particular, RTE-4), on real RTE data sets provided by the organizers of the challenges.

In order to highlight the phenomena relevant to component-based TE, we carried out a linguistically motivated analysis of entailment data. We presented a methodology for the creation of specialized TE data sets, made of atomic T-H pairs in which a certain phenomenon underlying the entailment relation is highlighted and isolated. Important outcomes of this thesis are the pilot resources obtained by the application of such methodology for the creation of specialized data sets, and by the application of the procedure for the acquisition of high precision entailment rules from

Wikipedia revision history. The first study resulted in the creation of two data sets, made of *i)* 90 RTE-5 Test Set pairs (30 entailment, 30 contradiction and 30 unknown examples) annotated with linguistic phenomena relevant to inference (both with fine grained and macro categories), and *ii)* 203 atomic pairs created from the 90 annotated pairs (157 entailment, 33 contradiction, and 13 unknown examples).
The results of the study on automatic knowledge acquisition, obtained on two experimental settings, respectively on causality (using the seed *because*) and temporal expressions (using the seed *before*) show both high quality and coverage of the extracted rules. The obtained resource includes, respectively, 1249 and 665 rules, which cover entailment and paraphrasing aspects not represented in other similar resources. Since the methodology does not require human intervention, the resource can be easily extended and periodically updated, as Wikipedia revisions change continuously. The resources described in this thesis, as well as new extensions, are freely available for research purposes on FBK HLT group website[1].

It seems premature to draw a definitive conclusion for our research on Component-Based TE, and efforts are still necessary in order to provide enough empirical evidence both in terms of the number of linguistic phenomena covered by the TE-components and in terms of the complexity and representativeness of the data sets used in the experiments. At the same time, we hope that the analysis of the different dimensions of the problem we provided may bring interesting elements to TE system developers to evaluate the potential impact of a solution to a specific sub-problem relevant to inference, and the interactions between linguistic phenomena. Evaluating the inference types a given system can reliably solve would make it easier to identify significant advances, and thereby promote the reuse of successful solutions and focus on unresolved aspects. For this rea-

[1] `http://hlt.fbk.eu/en/technology`

son we proposed an evaluation methodology to assess component-based TE systems capabilities to reliably solve sub-problems relevant to inference. Such methodology is based on the analysis of the system behaviour on atomic pairs with respect to the behaviour on corresponding original pairs. Through the definition of two indicators, a Component Correlation Index and a Component Deviation Index, we infer evaluation patterns which indicate strength and weaknesses of the system. As a pilot study we have applied our qualitative evaluation methodology to the output of three systems that took part in RTE-5, i.e. EDITS, VENSES (Venice Semantic Evaluation System) and BLUE (Boeing Language Understanding Engine), and we discovered that, although the three systems have similar accuracy on RTE-5 data sets, they show significant differences in their respective abilities to manage different linguistic phenomena and to properly combine them. As an outcome, a more meaningful evaluation of RTE systems is provided, that highlights on which aspects a system needs to improve its performance, and the features it should focus on.

The results obtained throughout our research work on Composition-Based TE and the interest showed by TE community towards this research direction encourage us to continue the investigation of this framework. We propose to exploit the results obtained in this research work to optimize specific TE component-based architectures for different applications (e.g. domain, genre), i.e. composed by modules that meet the requirements of that specific genre/domain. In line with this direction of domain-specific TE, we started an explorative study during the internship period at Xerox, whose main goal was to recognize Textual Entailment in Xerox Requests For Proposal responses (RFP) and contracts. Aim of the project was to support internal pre-sales services in the writing of responses to RFP, through the reuse of similar content contained in hand-crafted responses that have been written for previous customers. As a first step of

this study we collected and annotated a data set of 200 positive and negative text/hypothesis pairs from the database of past RFP responses and contracts, and we carried out preliminary evaluations.

Bibliography

[1] S. Ait-Mokhtar, J.P. Chanod, and C. Roux. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121144, 2002.

[2] E. Akhmatova and M. Dras. Using hypernymy acquisition to tackle (part of) textual entailment. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer 2009)*, Singapore. 6 August, 2009.

[3] I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, 2010.

[4] A. Balahur, E. Lloret, O. Ferrandez, A. Montoyo, M. Palomar, and R. Munoz. The dlsiuaes team's participation in the tac 2008 tracks. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland, USA. 17 November, 2008.

[5] R. Bar-Haim, J. Berant, I. Dagan, I. Greental, S. Mirkin, E. Shnarch, and I. Szpektor. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland, USA. 17 November, 2008.

[6] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy. 10 April, 2006.

[7] R. Bar-Haim, I. Dagan, and E. Shnarch. Semantic inference at the lexical-syntactic level. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, Vancouver, Canada. 22-26 July, 2007.

[8] R. Bar-Haim, I. Szpektor, and O. Glickman. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, MI. 30 June, 2005.

[9] R. Barzilay and K.R. McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–327, 2005.

[10] J. Van Benthem. The semantics of variety in categorial grammar. In W. Buszkowski, W. Marciszewski, and J. Van Benthem, editors, *Categorial Grammar*. John Benjamins, 1988.

[11] L. Bentivogli, E. Cabrio, I. Dagan, D. Giampiccolo, M. Lo Leggio, and B. Magnini. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. 19-21 May, 2010.

[12] L. Bentivogli, P. Clark, I. Dagan, H. T. Dang, and D. Giampiccolo. The sixth pascal recognizing textual entailment challenge. In *Pro-*

ceedings of the TAC 2009 Workshop on Textual Entailment, Gaithersburg, Maryland. 15-16 November, 2010.

[13] L. Bentivogli, I. Dagan, H.T. Dang, D. Giampiccolo, M. Lo Leggio, and B. Magnini. Considering discourse references in textual entailment annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, Pisa, Italy. 17-19 September, 2009.

[14] L. Bentivogli, B. Magnini, I. Dagan, H.T. Dang, and D. Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*, Gaithersburg, Maryland. 17 November, 2009.

[15] P. Blackburn, J. Bos, M. Kohlhase, and H. de Nivelle. Inference and computational semantics. *Studies in Linguistics and Philosophy, Computing Meaning*, 77(2):1128, 2001.

[16] J. Bos and K. Markert. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the second PASCAL Challenge Workshop on Recognizing Textual Entailment*, Venice, Italy. 10 April, 2006.

[17] A. Burchardt, M. Pennacchiotti, S. Thater, and M. Pinkal. Measures of the amount of ecologic association between species. *Natural Language Engineering (JNLE)*, 15(Special Issue 04), 2009.

[18] E. Cabrio, M. Kouylekov, and B. Magnini. Combining specialized entailment engines for rte-4. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, Gaithersburg, Maryland. 17-19 November, 2010.

[19] R. Carnap. *Logical Foundations of Probability 2nd ed.* Chicago, University of Chicago Press, 1962.

[20] N. Chambers, D. Cer, T. Grenager, D. Hall, C. Kiddon, B. MacCartney, D. Ramage, E. Yeh, and C.D. Manning. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. June 28-29, 2007.

[21] G. Chierchia and S. McConnell-Ginet. *Meaning and Grammar: An Introduction to Semantics 2nd ed.* Cambridge, MA: MIT Press, 2000.

[22] P. Clark and P. Harrison. Recognizing textual entailment with logical inference. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland, USA. 17 November, 2008.

[23] P. Clark and P. Harrison. An inference-based approach to recognizing entailment. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*, Gaithersburg, Maryland. 17 November, 2009.

[24] P. Clark, P. Harrison, J. Thompson, W. Murray, J. Hobbs, and C. Fellbaum. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. June 28-29, 2007.

[25] C. D. Manning D. Klein. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, Sapporo, Japan. 7-12 July, 2003.

[26] I. Dagan, R. Bar-Haim, I. Szpektor, I. Greental, and E. Shnarch. Natural language as the basis for meaning representation and inference. In *Proceedings of the 9th International Conference on Intel-*

ligent Text Processing and Computational Linguistics (CICLing08), Haifa, Israel. 17-23 February, 2008.

[27] I. Dagan, B. Dolan, B. Magnini, and D. Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, 15(Special Issue 04):i–xvii, 2009.

[28] I. Dagan and O. Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France. 26-29 January, 2004.

[29] I. Dagan, O. Glickman, and B. Magnini. The pascal recognizing textual entailment challenge. In *Proceedings of the First PASCAL Challenges Workshop on RTE*, Southampton, U.K. 12 April, 2005.

[30] I. Dagan, O. Glickman, and B. Magnini. The pascal recognizing textual entailment challenge. In *MLCW 2005, LNAI Volume 3944*. Springer-Verlag, 2006.

[31] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 3(7):171–176, 1964.

[32] R. Delmonte, A. Bristot, M.A. Piccolino Boniforti, and S. Tonelli. Entailment and anaphora resolution in rte-3. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. 28-29 June, 2007.

[33] R. Delmonte, S. Tonelli, and R. Tripodi. Semantic processing for text entailment with venses. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*, Gaithersburg, Maryland. 17 November, 2009.

[34] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[35] R. Cooper et al. Using the framework. In *Technical Report LRE 62-051 D-16, The FraCaS Consortium*, Prague, Czech Republic. June 28-29, 1996.

[36] C. Fellbaum. Wordnet: An electronic lexical database. In *Language, Speech and Communication*, MIT Press, 1998.

[37] G. Frege. Über sinn und bedeutung. In *Zeitschrift für Philosophie und philosophische Kritik*, volume 100.25-50, 1892.

[38] K. Garoufi. Towards a better understanding of applied textual entailment. In *Master Thesis*, Saarland University. Saarbrücken, Germany, 2007.

[39] D. Giampiccolo, H. Trang Dang, B. Magnini, I. Dagan, and E. Cabrio. The fourth pascal recognising textual entailment challenge. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland, USA. 17 November, 2008.

[40] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. The third pascal recognising textual entailment challenge. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. 28-29 June, 2007.

[41] C. Giuliano. jlsi a tool for latent semantic indexing. In *Software*, 2007.

[42] O. Glickman, I. Dagan, and M. Koppel. A lexical alignment model for probabilistic textual entailment. In *MLCW 2005, LNAI Volume 3944*. Springer-Verlag, 2006.

[43] S. Harabagiu, A. Hickl, and F. Lacatusu. Negation, contrast, and contradiction in text processing. In *Proceedings of the Twenty-First*

National Conference on Artificial Intellingence (AAAI-06), Boston, Massachusetts, 16-20 July, 2006.

[44] S. Harmeling. Inferring textual entailment with a probabilistically sound calculus. *Natural Language Engineering (JNLE)*, 15(Special Issue 04), 2009.

[45] R. Kirk. Building an annotated textual inference corpus for motion and space. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer 2009)*, Singapore. 6 August, 2009.

[46] J. Kittler, M. Hatef, R.P.W.Duin, and J. Mata. On combining classifiers. *IEEE Trans.*, 20, 1998.

[47] M. Kouylekov and B. Magnini. Tree edit distance for textual entailment. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RALNP-2005)*, Borovets, Bulgaria. 21-23 September, 2005.

[48] M. Kouylekov, Y. Mehdad, M. Negri, and E. Cabrio. Fbk participation in rte6: Main and kbp validation task. In *Proceedings of the Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland. 15-16 November, 2010.

[49] M. Kouylekov and M. Negri. An open-source package for recognizing textual entailment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) System Demonstrations*, Uppsala, Sweden. 11-16 July, 2010.

[50] Z. Kozareva and A. Montoyo. Mlent: The machine learning entailment system of the university of alicante. In *Proceedings of the second PASCAL Challenge Workshop on Recognizing Textual Entailment*, Venice, Italy. 10 April, 2006.

[51] D. Lin, , and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360, 2001.

[52] D. Lin. An information theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, San Francisco, California 24-27 July, 1998.

[53] B. MacCartney. Natural language inference. In *Phd dissertation*, June 2009.

[54] B. MacCartney and C.D. Manning. Natural logic for textual inference. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, Prague, Czech Republic. 23-30 June, 2007.

[55] B. MacCartney and C.D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK. 18-22 August, 2008.

[56] B. MacCartney and C.D. Manning. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8)*, Tilburg, The Netherlands. 7-9 January, 2009.

[57] B. Magnini and E. Cabrio. Combining specialized entailment engines. In *Proceedings of the 4th Language & Technology Conference (LTC'09)*, Poznan, Poland. 6-8 November, 2009.

[58] C.D. Manning. Local textual inference: its hard to circumscribe, but you know it when you see it - and nlp needs it. In *Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8)*, Unpublished manuscript. 25 February, 2006.

[59] M.C. De Marneffe, A.N. Rafferty, and C.D. Manning. Finding contradictions in text. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*, Columbus, OH. 15-20 June, 2008.

[60] A. Max and G. Wisniewski. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Valletta, Malta. 19-21 May, 2010.

[61] Y. Mehdad. Automatic cost estimation for tree edit distance using particle swarm optimization. In *In Proceedings of the 4the ACL-IJCNLP 2009 Conference*, Singapore. 2-7 August, 2009.

[62] Y. Mehdad and B. Magnini. Optimizing textual entailment using particle swarm optimization. In *In Proceedings of the TextInfer Workshop*, Singapore. 2-7 August, 2009.

[63] Y. Mehdad and B. Magnini. A word overlap baseline for the recognizing textual entailment task. In *Available at*, 2009.

[64] Y. Mehdad, M. Negri, E. Cabrio, M. Kouylekov, and B. Magnini. Using lexical resources in a distance-based approach to rte. In *Proceedings of the TAC 2009 Workshop on TE*, Gaithersburg, Maryland. 17 November, 2009.

[65] S. Mirkin, R. Bar-Haim, J. Beran, I. Dagan, E. Shnarch, A. Stern, and I. Szpektor. Addressing discourse and document structure in the rte search task. In *Proceedings of the TAC 2009 Workshop on TE*, Gaithersburg, Maryland. 17 November, 2009.

[66] S. Mirkin, J. Berant, I. Dagan, and Eyal Shnarch. Recognising entailment within discourse. In *Proceedings of the 23rd International*

Conference on Computational Linguistics (COLING 2010), Beijing, China. 23-27 August, 2010.

[67] S. Mirkin, I. Dagan, and Sebastian Padò. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden. 11-16 July, 2010.

[68] S. Mirkin, L. Specia, N. Cancedda, I. Dagan, M. Dymetman, and I. Szpektor. Source-language entailment modeling for translating unknown terms. In *In Proceedings of the 47th Annual Meeting of ACL and the 4th International Joint Conf. on Natural Language Processing of AFNLP*, Singapore. 2-7 August, 2009.

[69] J. Mitchell and M.Lapata. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*, Columbus, OH. 15-20 June, 2008.

[70] C. Monz and M. de Rijke. Light-weight entailment checking for computational semantics. In *Proceedings Inference in Computational Semantics (ICoS-3)*, pages 59–72, 2001.

[71] R. Nairn, C. Condoravdi, and L. Karttunen. Computing relative polarity for textual inference. In *Inference in Computational Semantics (ICoS-5)*, Buxton, UK. 20-21 April, 2006.

[72] M. Negri, M. Kouylekov, B. Magnini, Y. Mehdad, and E. Cabrio. Towards extensible textual entailment engines: The edits package. In *AI*IA 2009: Emergent Perspectives in Artificial Intelligence, Lecture Notes in Computer Science. Volume 5883*, Springer-Verlag Berlin Heidelberg, p. 314., 2009.

[73] A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Computational Logic*, V, No. N, February:1–23, 2007.

[74] R. D. Nielsen, W. Ward, and J. H. Martin. Recognizing entailment in intelligent tutoring systems. *The Journal of Natural Language Engineering, (JNLE)*, 15:479–501, 2009.

[75] J. Nolt, D. Rohatyn, and A. VArzi. *Schaum's outline of Theory and Problems of Logic 2nd ed.* McGraw-Hill, 1998.

[76] S. Padò, M. Galley, D. Jurafsky, and C.D. Manning. Robust machine translation evaluation with entailment features. In *In Proceedings of the 47th Annual Meeting of ACL and the 4th International Joint Conf. on Natural Language Processing of AFNLP*, Singapore. 2-7 August, 2009.

[77] S. Padò and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 31(1):71–105, 2007.

[78] Ted Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, CA. 25-29 July, 2004.

[79] M. Pinkal. Logic and lexicon: the semantics of the indenite. *Studies in linguistics and philosophy*, 56, 1995.

[80] I. Dagan. R. B. Aharon, I. Szpektor. Generating entailment rules from framenet. In *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden. July 11-16, 2010.

[81] L. Romano, M. O. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy. 3-7 April, 2006.

[82] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall Series in Artificial Intelligence, 2002.

[83] M. Sammons, V.G.V Vydiswaran, and D. Roth. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden. 11-16 July, 2010.

[84] S. Sekine. Automatic paraphrase discovery based on context and kwywords between ne pairs. In *Proceedings of the International Workshop on Paraphrasing (IWP-05)*, Jeju Island, South Korea. 14 October, 2005.

[85] R. Siblini and L. Kosseim. Using ontology alignment for the tac rte challenge. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland. 17 November, 2008.

[86] I. Szpektor, E. Shnarch, and I Dagan I. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, Prague, Czech Republic. 23-30 June, 2007.

[87] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling web-based acquisition of entailment relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. 25-26 July, 2004.

[88] M. Tatu and D. Moldovan. Cogex at rte3. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. 23-30 June, 2007.

[89] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated dp based search for statistical translation. In *Proceedings of the European Conf. on Speech Communication and Technology*, September. Rhodes, Greece, 1997.

[90] V. Sánchez Valencia. Studies on natural logic and categorial grammar. In *PhD Thesis, University of Amsterdam*, Uppsala, Sweden. 11-16 July, 1991.

[91] L. Vanderwende, D. Coughlin, and B. Dolan. What syntax can contribute in entailment task. In *Proceedings of the First PASCAL Challenges Workshop on RTE*, Southampton, U.K., 11-13 April, 2005.

[92] L. Vanderwende, A. Menezes, and R. Snow. Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy. 10 April, 2006.

[93] D. Walton and C. A. Reed. Argumentation schemes and enthymemes. *Synthese*, 145:339–370, 2005.

[94] R. Wang and G. Neumann. An accuracy-oriented divide-and-conquer strategy. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland. 17 November, 2008.

[95] R. Wang and Y. Zhang. Recognizing textual entailment with temporal expressions in natural language texts. In *Proceedings of the IEEE International Workshop on Semantic Computing and Applications (IWSCA-2008)*, Incheon, South Korea. 10-11 July, 2008.

[96] R. Wang, Y. Zhang, and G. Neumann. A joint syntactic-semantic representation for recognizing textual relatedness. In *Proceedings of the TAC 2009 Workshop on TE*, Gaithersburg, Maryland. 17 November, 2009.

[97] M.A. Yatbaz. Rte4: Normalized dependency tree alignment using unsupervised n-gram word similarity score. In *Proceedings of the TAC 2008 Workshop on Textual Entailment*, Gaithersburg, Maryland, USA. 17 November, 2008.

[98] A. Zaenen, L. Karttunen, and R. Crouch. Local textual inference: can it be dened or circumscribed? In *Proceedings of the Workshop on the Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, MI. 30 June, 2005.

[99] F.M. Zanzotto and M. Pennacchiotti. Expanding textual entailment corpora from wikipedia using co-training. In *Proceedings of the COLING-Workshop on The Peoples Web Meets NLP: Collaboratively Constructed Semantic Resources*, Beijing, China. 28 August, 2010.

[100] F.M. Zanzotto, M. Pennacchiotti, and A. Moschitti. Shallow semantics in fast textual entailment rule learners. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. 23-30 June, 2007.

[101] F.M. Zanzotto, M. Pennacchiotti, and A. Moschitti. A machine learning approach to textual entailment recognition. *Natural Language Engineering (JNLE)*, 15(Special Issue 04), 2009.

[102] K. Zhang and D. Shasha. Fast algorithm for the unit cost editing distance between trees. *Journal of Algorithms*, 11, December 1990.

Appendix A

List of Published Papers

2011

Elena Cabrio, Bernardo Magnini, *Defining Specialized Entailment Engines Using Natural Logic Relations*, To appear in: Zygmunt Vetulani, LTC2009 Revised Selected Papers. Lecture Notes in Artificial Intelligence. Volume 6562.

ABSTRACT: In this paper we propose a framework for the definition and combination of *specialized entailment engines*, each of which able to deal with a certain aspect of language variability. Such engines are based on transformations, and we define them taking advantage of the conceptual and formal tools available from an extended model of Natural Logic (NL). Given a T,H pair, each engine performs atomic edits to solve the specific linguistic phenomenon it is built to deal with, and assigns an entailment relation as the output of this operation. NL mechanisms of semantic relations composition are then applied to join the output of each single engine, in order to obtain a global entailment judgement for a pair.

Elena Cabrio, Bernardo Magnini, *Towards Component-Based Textual Entailment*, Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), Oxford, UK, January 12-14, 2011.

ABSTRACT: In the Textual Entailment community, a shared effort towards a deeper understanding of the core phenomena involved in textual inference is recently arose. To analyse how the common intuition that decomposing TE would allow a better comprehension of the problem from both a linguistic and a computational viewpoint, we propose a definition for *strong component-based TE*, where each component is in itself a complete

TE system, able to address a TE task on a specific phenomenon in isolation. We review the literature according to our definition, trying to position relevant work as more or less close to our idea of strong component-based TE. Several dimensions of the problem are discussed: *i)* the implementation of system components to address specific inference types, *ii)* the analysis of the phenomena relevant to component-based TE, and *iii)* the development of evaluation methodologies to assess TE systems capabilities to address single phenomena in a pair (`http://www.aclweb.org/anthology/sigsem.html#2011_0`).

2010

Milen Kouylekov, Yashar Mehdad, Matteo Negri, Elena Cabrio, *FBK Participation in RTE6: Main and KBP Validation Task*, Proceedings of the Text Analysis Conference (TAC 2010), Gaithersburg, Maryland, USA, November 15-16, 2010.

ABSTRACT: This paper overviews FBK's participation in the Main and KBP Validation Pilot task organized within the RTE6 Evaluation Campaign. Our submissions have been produced running the EDITS (Edit Distance Textual Entailment Suite) open source RTE package, which allows to experiment with different combinations of algorithms, entailment rules, and optimization strategies. The evaluation on test data confirmed their effectiveness, with good results in both the tasks. Our best run in the Main task achieved a Micro-Averaged F-measure of 44.71% (with the best and the median system respectively achieving 48.01% and 33.72%); our best run in the KBP Validation task achieved the highest score, with 25.5% F-measure (it will be available here: `http://www.nist.gov/tac/publications/index.html`)

Elena Cabrio, Bernardo Magnini, *Toward Qualitative Evaluation of Textual Entailment Systems*, Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010: Poster), Beijing, China, August 23-27, 2010.

ABSTRACT: This paper presents a methodology for a quantitative and qualitative evaluation of Textual Entailment systems. We take advantage of the decomposition of Text Hypothesis pairs into *monothematic pairs*, i.e. pairs where only one linguistic phenomenon at a time is responsible for entailment judgement, and propose to run TE systems over such datasets. We show that several behaviours of a system can be explained in terms of the correlation between the accuracy on monothematic pairs and the accuracy on the corresponding original pairs (`www.aclweb.org/anthology/C/C10/C10-2000.pdf`).

Bernardo Magnini, Elena Cabrio, *Contradiction-Focused Qualitative Evaluation of Textual Entailment*, Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (Ne-Sp NLP 2010), Uppsala, Sweden, July 10, 2010.

ABSTRACT: In this paper we investigate the relation between positive and negative pairs in Textual Entailment (TE), in order to highlight the role of contradiction in TE datasets. We base our analysis on the decomposition of Text-Hypothesis pairs into *monothematic pairs*, i.e. pairs where only one linguistic phenomenon at a time is responsible for entailment judgement and we argue that such a deeper inspection of the linguistic phenomena behind textual entailment is necessary in order to highlight the role of contradiction. We support our analysis with a number of empirical experiments, which use current available TE systems (`portal.acm.org/citation.cfm?id=1858973`).

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, Bernardo Magnini, *Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference*, in Proceedings of the Language Resources and Evaluation Conference (LREC 2010), Malta, May 19-21, 2010.

ABSTRACT: This paper proposes a methodology for the creation of specialized data sets for Textual Entailment, made of monothematic Text-Hypothesis pairs (i.e. pairs in which only one linguistic phenomenon relevant to the entailment relation is highlighted and isolated). The expected benefits derive from the intuition that investigating the linguistic phenomena separately, i.e. decomposing the complexity of the TE problem, would yield an improvement in the development of specific strategies to cope with them. The annotation procedure assumes that humans have knowledge about the linguistic phenomena relevant to inference, and a classification of such phenomena both into fine grained and macro categories is suggested. We experimented with the proposed methodology over a sample of pairs taken from the RTE-5 data set, and investigated critical issues arising when entailment, contradiction or unknown pairs are considered. The result is a new resource, which can be profitably used both to advance the comprehension of the linguistic phenomena relevant to entailment judgements and to make a first step towards the creation of large-scale specialized data sets (`http://hlt.fbk.eu/en/Technology/TE_Specialized_Data`).

2009

Elena Cabrio, Yashar Mehdad, Matteo Negri, Milen Kouylekov, Bernardo Magnini, *Recognizing Textual Entailment for Italian: EDITS@EVALITA 2009*, in Proceedings of AI*IA 2009, Reggio Emilia, Italy, December 9-12, 2009.

ABSTRACT: This paper overviews FBK's participation in the Textual Entailment task at EVALITA 2009. Our runs were obtained through different configurations of EDITS (Edit Distance Textual Entailment Suite), the first freely available open source tool for Recognizing Textual Entailment (RTE). With a 71% Accuracy, EDITS reported the best score out of the 8 submitted runs. We describe the sources of knowledge that have been used (e.g. extraction of rules from Wikipedia), the different algorithms applied (i.e. Token Edit Distance, Tree Edit Distance), and the Particle Swarm Optimization (PSO) module used to estimate the optimal cost of edit operations in the cost scheme. Two different dependency parsers for the annotation of the data in the preprocessing phase have been compared, to assess the impact of the parser on EDITS performances. Finally, the obtained results and error analysis are discussed (`evalita.fbk.eu/reports/Textual\%20Entailment/TE_FBK_UNITN.pdf`).

Matteo Negri, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, Elena Cabrio, *Towards Extensible Textual Entailment Engines: the EDITS Package*, AI*IA 2009: Emergent Perspectives in Artificial Intelligence, Lecture Notes in Computer Science, 2009, Volume 5883/2009.

ABSTRACT: This paper presents the first release of EDITS, an open-source software package for recognizing Textual Entailment developed by FBK-irst. The main contributions of EDITS consist in: *i)* providing a basic framework for a distance-based approach to the task, *ii)* providing a highly customizable environment to experiment with different algorithms, *iii)* allowing for easy extensions and integrations with new algorithms and resources. System's main features are described, together with experiments over different datasets showing its potential in terms of tuning and adaptation capabilities (`http://www.springerlink.com/content/a331554882218573/`).

Yashar Mehdad, Matteo Negri, Elena Cabrio, Milen Kouylekov, Bernardo Magnini, *Using Lexical Resources In a Distance-Based Approach to RTE*, in Proceedings of the Text Analysis Conference (TAC 2009), Gaithersburg, Maryland, USA, November 17, 2009.

ABSTRACT: This paper overviews FBK's participation in the RTE 5 Evaluation Campaign. Our runs, submitted both to the main (two-way classification), and to the pilot

task, were obtained through different configurations of EDITS (Edit Distance Textual Entailment Suite) package, the first freely available open source RTE software. The main sources of knowledge used, the different configurations, and the achieved results are described, together with ablation tests representing a preliminary analysis of the actual contribution of different resources to the RTE task (`http://www.nist.gov/tac/publications/2009/papers.html`).

Bernardo Magnini, Elena Cabrio, *Combining Specialized Entailment Engines*, in Proceedings of the 14th Language and technology conference (LTC'09), Poznan, Poland, November 6-8, 2009.
ABSTRACT: In this paper we propose a general method for the combination of specialized textual entailment engines. Each engine is supposed to address a specific language phenomenon, which is considered relevant for drawing semantic inferences. The model is based on the idea that the distance between the Text and the Hypothesis can be conveniently decomposed into a combination of distances estimated by single and disjoint engines over distinct linguistic phenomena. We provide both the formal definition of the model and preliminary empirical evidences supporting the underlying intuition).

Elena Cabrio, *Specialized Entailment Engines: Approaching Linguistics Aspects of Textual Entailment*, in Helmut Horacek, Elisabeth Métais, Rafael Muñoz, Magdalena Wolsks (Eds.), *Natural Language Processing and Information Systems*, Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB 2009), Saarbruecken, Germany, June 24-26, 2009. Springer LNCS, Lecture Notes in Computer Science, 2010, Volume 5723/2010. Best Paper Award at the Doctoral Symposium.
ABSTRACT: Textual Entailment (TE), one of the current hot topics in Computational Linguistics, has been proposed as a task to address the problem of language variability. Since TE is due to the combination of different linguistic phenomena which interact among them in a complex way, this paper proposes to experiment the use of specialized entailment engines, each addressing a specific phenomenon relevant to entailment (`http://www.springerlink.com/content/p28n5v6843p88415/`).

2008

Elena Cabrio, Milen Kouylekov, Bernardo Magnini, *Combining Special-*

***ized Entailment Engines for RTE-4*, in Proceedings of the First Text Analysis Conference (TAC 2008), Gaithersburg, Maryland, USA, November 17-19, 2008.**

ABSTRACT: The main goal of FBK-irst participation at RTE-4 was to experiment the use of combined specialized entailment engines, each addressing a specific phenomena relevant to entailment. The approach is motivated since textual entailment is due to the combination of several linguistic phenomena which interact among them in a quite complex way. We were driven by the following two considerations: *(i)* devise a general framework, based on distance between T and H, flexible enough to allow the combination of single entailment engines; *(ii)* provide a modular approach through which evaluate progresses on single aspects of entailment, using specialised training and test dataset. For RTE-4 we used two simple entailment engines, one addressing negation and the other lexical similarity, with a linear combination of their respective distances on T-H pairs (`http://www.nist.gov/tac/publications/2008/papers.html`).

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, Bill Dolan, *The Fourth PASCAL Recognizing Textual Entailment Challenge*, in Proceedings of the First Text Analysis Conference (TAC 2008), Gaithersburg, Maryland, USA, November 17-19, 2008.

ABSTRACT: In 2008 the Recognizing Textual Entailment Challenge (RTE-4) was proposed for the first time as a track at the Text Analysis Conference (TAC). Another important innovation introduced in this campaign was a three-judgement task, which required the systems to make a further distinction between the pairs where the entailment does not hold because the content of H is contradicted by the content of T and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T. A classic two-way task was also offered. RTE-4 attracted 26 teams, more than an half of whom submitted runs for the new 3-way task. This paper describes the preparation of the data set, and gives an overview of the results achieved by the participating systems (`http://www.nist.gov/tac/publications/2008/papers.html`).