Chapter 1

Introduction

This thesis is about the meaning relations that can hold between language expressions (words, phrases, clauses, and sentences). In particular, it focuses on the meaning relations of paraphrasing, textual entailment, and semantic similarity. The automatic processing of these meaning relations is an unsolved problem in Computational Linguistics (CL) and Natural Language Processing (NLP) and has attracted the attention of many researchers. This thesis explores two different directions within the research on meaning relations:

1. Incorporating linguistic knowledge in the empirical tasks of processing meaning relations. In particular, I focus on the paraphrasing meaning relation and the empirical task of Paraphrase Identification (PI). By combining PI with Paraphrase Typology (PT) I aim:

    a) to improve the evaluation and interpretation of automated PI systems.

    b) to empirically validate PT.

2. Analyzing and processing multiple meaning relations together. I contrast previous work and propose a novel research approach that does not focus on a single meaning relation. I present a joint study on Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity:

    a) to compare the different meaning relations empirically.

    b) to create a shared typology for textual meaning relations.

My work offers a valuable insight into the nature and interactions of the different meaning relations and also aims to improve the automated systems for processing meaning relations. I also release to the community three new corpora, two new typologies of meaning relations, a new web-based annotation tool, and a new

software program for a qualitative evaluation of automated paraphrase identification systems.

The structure of this thesis is intentionally chronological[1] in order to capture the four year development of the ideas and arguments behind the thesis. The thesis consists of nine Chapters, organized as follows:

- Chapter 1 is the Introduction.
- Chapters 2 to 8 correspond to seven published articles. They are grouped in three thematically organized parts.
- Chapter 9 presents the contributions, the discussion of the results, the conclusions, and the directions for future work.

The rest of this Introduction chapter is organized as follows. In Section 1.1, I familiarize readers with the related work in the research on meaning relations. In Section 1.2, I present my main objectives and justify them in the context of the preexisting research. In Section 1.3, I describe the development of this thesis and the connecting thread that runs between the individual articles. Finally in Section 1.4, I present the outline and structure of the whole dissertation.

1.1 Related Work

This section is meant to provide the reader with a compact overview of the previous and latest research related to this thesis in order to supplement and bind together the "background" sections in each paper. From a thematic perspective, the subject matter can be broken down into the following research areas:

(i) Textual Meaning Relations. Empirical Tasks. (**Section 1.1.1**)

(ii) Typologies of Textual Meaning Relations (**Section 1.1.2**)

(iii) Joint Research on Textual Meaning Relations (**Section 1.1.3**)

(iv) Other Related Work (**Section 1.1.4**)

I will deal with each of the areas in turn, highlighting the main trends and milestones. The reader is referred to the original papers for details.

[1] Articles are presented in the order in which they were written, which does not necessarily correspond to the order in which they were published.

1.1.1 Textual Meaning Relations. Empirical Tasks

Meaning relations between complex language expressions (e.g.: clauses, sentences, paragraphs), henceforth "textual meaning relations" are the object of study of this thesis. Research on textual meaning relations has to account not only for the meaning of a single word or a phrase, but also for the compositionality of meaning. In this thesis, I focus on the textual meaning relations of Paraphrasing, Textual Entailment, Contradiction[2], and Semantic Similarity. It is important to note that the interactions between the different relations are non-trivial. In some cases they can overlap (e.g.: two texts that are paraphrases often also hold an entailment relation) and in some cases the negative examples for one relation can be positive examples for another (e.g.: two texts that are not paraphrases can sometimes hold an entailment relation or a contradiction relation).

Empirical Tasks on Textual Meaning Relations
Androutsopoulos and Malakasiotis [2010] distinguish three types of empirical tasks that are focused on processing meaning relations: recognition, generation, and extraction. Their definitions for these paraphrasing and textual entailment tasks are as follows:

> **Recognition**: *"The main input to a paraphrase or textual entailment recognizer is a pair of language expressions (or templates), possibly in particular context. The output is a judgment, possibly probabilistic, indicating whether or not the members of the input pair are paraphrases or a correct textual entailment pair; the judgments must agree as much as possible with those of humans."*
>
> **Generation**: *"The main input to a paraphrase or textual entailment generator is a single language expression (or template) at a time, possibly in a particular context. The output is a set of paraphrases of the input or a set of language expressions that entail or are entailed by the input; the output set must be as large as possible, but including as few errors as possible."*
>
> **Extraction**: *"The main input to a paraphrase or textual entailment extractor is a corpus, for example a monolingual corpus of parallel or comparable texts. The system outputs pairs of paraphrases (possibly templates) or pairs of language expressions (or templates) that constitute correct textual entailment pairs, based on the evidence of the corpus; the goal is again to produce as many output pairs as possible, with as few errors as possible."*

[2]Contradiction is typically studied jointly with Textual Entailment.