PARAPHRASING, TEXTUAL ENTAILMENT, AND SEMANTIC SIMILARITY ABOVE WORD LEVEL

VENELIN ORLINOV KOVATCHEV

Tesis presentada para optar
al grado de **Doctor en Linguística** con mención europea
en el programa de doctorado *Ciencia Cognitiva y Lenguaje*,
Departament de Filologia Catalana i Lingüística General,
Universidad de Barcelona,

bajo la supervisión de

Dra. M. Antònia Martí
Universidad de Barcelona

Dra. Maria Salamó
Universidad de Barcelona

Mayo de 2020

To Mila, who supported me every step of the way.

To Maya and Orlin, for always encouraging my curiosity.

Abstract

This dissertation explores the linguistic and computational aspects of the meaning relations that can hold between two or more complex linguistic expressions (phrases, clauses, sentences, paragraphs). In particular, it focuses on Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity. This thesis is composed of seven different articles and is divided into three thematic Parts.

In *Part I: "Similarity at the Level of Words and Phrases"*, I study the Distributional Hypothesis (DH). DH is central for most contemporary approaches for automatic processing of meaning and meaning relations within Computational Linguistics (CL) and Natural Language Processing (NLP). Part I of this thesis explores different methodologies for quantifying semantic similarity at the levels of words and short phrases. I measure the importance of the corpus size and the role of linguistic preprocessing. I also show that (lexical) semantic similarity can interact with syntactic-based compositional rules and result in productive patterns at the phrase level. The research in Part I resulted in the publication of two articles.

In *Part II: "Paraphrase Typology and Paraphrase Identification"*, I focus on the meaning relation of paraphrasing and the empirical task of automated Paraphrase Identification (PI). Paraphrasing is one of the most widely studied meaning relation both in theoretical and practical research. PI is among the most popular tasks in CL and NLP. In Part II of this thesis I present: 1) EPT: a new typology of the linguistic and reason-based phenomena involved in paraphrasing; 2) WARP-Text: a new web-based annotation interface capable of annotating paraphrase types; 3) ETPC: the largest corpus to date to be annotated with paraphrase types; and 4) a qualitative evaluation framework for automated PI systems. The findings presented in Part II provide in-depth knowledge on the nature of the paraphrasing relation and improve the evaluation, interpretation, and error analysis in the task of PI. The research in Part II resulted in the publication of three articles.

In *Part III:"Paraphrasing, Textual Entailment, and Semantic Similarity"*, I present a novel direction in the research on textual meaning relations, resulting from joint research carried out on on paraphrasing, textual entailment, contradiction, and semantic similarity. Traditionally, these meaning relations are studied in isolation and the transfer of knowledge and resources between them is limited.