Chapter 5

ETPC - a Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó
University of Barcelona

Published at
Proceedings of the
Eleventh International Conference on Language Resources and Evaluation, 2018
pp.: 1384-1392

Abstract We present the Extended Paraphrase Typology (EPT) and the Extended Typology Paraphrase Corpus (ETPC). The EPT typology addresses several practical limitations of existing paraphrase typologies: it is the first typology that copes with the non-paraphrase pairs in the paraphrase identification corpora and distinguishes between contextual and habitual paraphrase types. ETPC is the largest corpus to date annotated with atomic paraphrase types. It is the first corpus with detailed annotation of both the paraphrase and the non-paraphrase pairs and the first corpus annotated with paraphrase and negation. Both new resources contribute to better understanding the paraphrase phenomenon, and allow for studying the relationship between paraphrasing and negation. To the developers of Paraphrase Identification systems ETPC corpus offers better means for evaluation and error analysis. Furthermore, the EPT typology and ETPC corpus emphasize the relationship with other areas of NLP such as Semantic Similarity, Textual Entailment, Summarization and Simplification.

Keywords Paraphrasing, Paraphrase Typology, Paraphrase Identification

5.1 Introduction

The task of Paraphrase Identification (PI) consists of comparing two texts of arbitrary size in order to determine whether they have approximately the same meaning. The most common approach to PI is as a binary classification problem, in which a system learns to make correct binary predictions (paraphrase or non-paraphrase) for a given pair of texts. The task of PI is challenging from more than one point of view. From the resource point of view, defining the task and preparing high quality corpora is a non-trivial problem due to the complex nature of "paraphrasing". From the application point of view, for a system to perform well on PI often requires a complex ML architecture and/or a large set of manually engineered features. From the evaluation point of view, the classical task of PI does not offer many possibilities for detailed error analysis, which in turn limits the reusability and the improvement of PI systems.

In the last few years, researchers in the field of paraphrasing have adopted the approach of decomposing the meta phenomenon of *"textual paraphrasing"* into a set of *"atomic paraphrase"* phenomena, which are more strictly defined and easier to work with. *"Atomic paraphrases"* are hierarchically organized into a typology, which provides a better means to study and understand paraphrasing. While the theoretical advantages of these approaches are clear, their practical implications have not been fully explored. The existing corpora annotated with paraphrase typology are limited in size, coverage and overall quality. The only corpus of sufficient size to date annotated with paraphrase typology is the corpus by Vila et al. [2015], which contains 3900 re-annotated *"textual paraphrase"* pairs from the MRPC corpus [Dolan et al., 2004].

The use of a paraphrase typology in practical tasks has several advantages. First, *"atomic paraphrases"* are much more strict in their definition, which makes the results more useful and understandable. Second, the more detailed annotation can be useful to (re)balance binary PI corpora in terms of type distribution. Third, annotating a corpus with paraphrase types provides much better feedback to the PI systems and allows for a detailed, per-type error analysis. Fourth, enriching the corpus and improving the evaluation can provide a linguistic insight into the workings of complex machine learning systems (i.e. Deep Learning) that are traditionally hard to interpret. Fifth, corpora annotated with a paraphrase typology open the way for new research and new tasks, such as "PI by type“ or "Atomic PI in context". Finally, decomposing *"textual paraphrases"* can help relate the task of PI to tasks such as Recognizing Textual Entailment, Text Summarization, Text Simplification, and Question Answering.