Chapter 6

A Qualitative Evaluation Framework for Paraphrase Identification

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

Published at
Proceedings of the
Twelfth Recent Advances in Natural Language Processing Conference, 2019
pp.: 569-579

Abstract In this paper, we present a new approach for the evaluation, error analysis, and interpretation of supervised and unsupervised Paraphrase Identification (PI) systems. Our evaluation framework makes use of a PI corpus annotated with linguistic phenomena to provide a better understanding and interpretation of the performance of various PI systems. Our approach allows for a qualitative evaluation and comparison of the PI models using human interpretable categories. It does not require modification of the training objective of the systems and does not place additional burden on the developers. We replicate several popular supervised and unsupervised PI systems. Using our evaluation framework we show that: 1) Each system performs differently with respect to a set of linguistic phenomena and makes qualitatively different kinds of errors; 2) Some linguistic phenomena are more challenging than others across all systems.

6.1 Introduction

In this paper we propose a new approach to evaluation, error analysis and interpretation in the task of Paraphrase Identification (PI). Typically, PI is defined as comparing two texts of arbitrary size in order to determine whether they have approximately the same meaning [Dolan et al., 2004]. The two texts in 1a and 1b are considered paraphrases, while the two texts at 2a and 2b are non-paraphrases.[1] In 1a and 1b there is a change in the wording (*"magistrate"* - *"judge"*) and the syntactic structure (*"was ordered"* - *"ordered"*) but the meaning of the sentences is unchanged. In 2a and 2b there are significant differences in the quantities (*"5%"* - *"4.7%"* and *"$27.45"* - *"$27.54"*).

1a A federal magistrate in Fort Lauderdale ordered him held without bail.

1b He was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.

2a Microsoft fell **5 percent** before the open to **$27.45** from Thursday's close of $28.91.

2b Shares in Microsoft slipped **4.7 percent** in after-hours trade to **$27.54** from a Nasdaq close of $28.91.

The task of PI can be framed as a binary classification problem. The performance of the different PI systems is reported using the Accuracy and F1 score measures. However this form of evaluation does not facilitate the interpretation and error analysis of the participating systems. Given the Deep Learning nature of most of the state-of-the-art systems and the complexity of the PI task, we argue that better means for evaluation, interpretation, and error analysis are needed. We propose a new evaluation methodology to address this gap in the field. We demonstrate our methodology on the ETPC corpus [Kovatchev et al., 2018a] - a recently published corpus, annotated with detailed linguistic phenomena involved in paraphrasing.

We replicate several popular state-of-the-art Supervised and Unsupervised PI Systems and demonstrate the advantages of our evaluation methodology by analyzing and comparing their performance. We show that while the systems obtain similar quantitative results (Accuracy and F1), they perform differently with respect to a set of human interpretable linguistic categories and make qualitatively different kinds of errors. We also show that some of the categories are more challenging than others across all evaluated systems.

[1] Examples are from the MRPC corpus [Dolan et al., 2004]

6.2 Related Work

The systems that compete on PI range from using hand-crafted features and Machine Learning algorithms [Fernando and Stevenson, 2008, Madnani et al., 2012, Ji and Eisenstein, 2013] to end-to-end Deep Learning models [He et al., 2015, He and Lin, 2016, Wang et al., 2016, Lan and Xu, 2018b, Kiros et al., 2015, Conneau et al., 2017]. The PI systems are typically divided in two groups: Supervised PI systems and Unsupervised PI systems.

"Supervised PI systems" [He et al., 2015, He and Lin, 2016, Wang et al., 2016, Lan and Xu, 2018b] are explicitly trained for the PI task on a PI corpora. "Unsupervised PI systems" in the PI field is a term used for systems that use a general purpose sentence representations such as Mikolov et al. [2013b], Pennington et al. [2014], Kiros et al. [2015], Conneau et al. [2017]. To predict the paraphrasing relation, they can compare the sentence representations of the candidate paraphrases directly (ex.: cosine of the angle), and use a PI corpus to learn a threshold. Alternatively they can use the representations as features in a classifier.

The complexity of paraphrasing has been emphasized by many researchers [Bhagat and Hovy, 2013, Vila et al., 2014, Benikova and Zesch, 2017]. Similar observations have been made for Textual Entailment [Sammons et al., 2010, Cabrio and Magnini, 2014]. Gold et al. [2019] study the interactions between paraphrasing and entailment.

Despite the complexity of the phenomena, the popular PI corpora [Dolan et al., 2004, Ganitkevitch et al., 2013, Iyer et al., 2017, Lan et al., 2017] are annotated in a binary manner. In part it is due to lack of annotation tools capable of fine-grained annotation of relations. WARP-Text [Kovatchev et al., 2018b] fills this gap in the NLP toolbox.

The simplified corpus format poses a problem with respect to the quality of the PI task and the ways it can be evaluated. The vast majority of the state-of-the-art systems in PI provide no or very little error analysis. This makes it difficult to interpret the actual capabilities of a system and its applicability to other corpora and tasks.

Some researchers have approached the problem of non-interpretability by evaluating the same architecture on multiple datasets and multiple tasks. Lan and Xu [2018a] apply this approach to Supervised PI systems, while Aldarmaki and Diab [2018] use it for evaluating Unsupervised PI systems and general sentence representation models.

Linzen et al. [2016] demonstrate how by modifying the task definition and the evaluation the capabilities of a Deep Learning system can be determined implicitly. The main advantage of such an approach is that it only requires modification and additional annotation of the corpus. It does not place any additional burden on the developers of the systems and can be applied to multiple systems without