Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 59–73

www.elsevier.com/locate/csl

Combining sentence similarities measures to identify paraphrases[☆]

Rafael Ferreira[*,a], George D.C. Cavalcanti[b], Fred Freitas[b], Rafael Dueire Lins[a], Steven J. Simske[c], Marcelo Riss[d]

[a] *Department of Statistics and Informatics, Federal Rural University of Pernambuco, Recife, Pernambuco, Brazil*
[b] *Informatics Center, Federal University of Pernambuco, Recife, Pernambuco, Brazil*
[c] *Hewlett-Packard, Fort Collins, CO 80528, USA*
[d] *Hewlett-Packard Brazil, Porto Alegre, Rio Grande do Sul, Brazil*

Received 7 June 2016; received in revised form 13 January 2017; accepted 1 July 2017
Available online 11 July 2017

Abstract

Paraphrase identification consists in the process of verifying if two sentences are semantically equivalent or not. It is applied in many natural language tasks, such as text summarization, information retrieval, text categorization, and machine translation. In general, methods for assessing paraphrase identification perform three steps. First, they represent sentences as vectors using bag of words or syntactic information of the words present the sentence. Next, this representation is used to measure different similarities between two sentences. In the third step, these similarities are given as input to a machine learning algorithm that classifies these two sentences as paraphrase or not. However, two important problems in the area of paraphrase identification are not handled: (i) the meaning problem: two sentences sharing the same meaning, composed of different words; and (ii) the word order problem: the order of the words in the sentences may change the meaning of the text. This paper proposes a paraphrase identification system that represents each pair of sentence as a combination of different similarity measures. These measures extract lexical, syntactic and semantic components of the sentences encompassed in a graph. The proposed method was benchmarked using the Microsoft Paraphrase Corpus, which is the publicly available standard dataset for the task. Different machine learning algorithms were applied to classify a sentence pair as paraphrase or not. The results show that the proposed method outperforms state-of-the-art systems.

Keywords: Sentence similarity; Paraphrase identification; Sentence simplification; Graph-based model

1. Introduction

The degree of similarity between phrases is measured by sentence similarity, or short-text similarity methods. These similarity methods should also address problems of measuring sentences with partial information, such as when one sentence is split into two or more short texts or phrases that contain two or more sentences. One specific task derived from sentence similarity is the Paraphrase Identification (PI). This task aims to verify if two sentences

[☆] This paper has been recommended for acceptance by Pascale fung.

[*] Corresponding author.

E-mail addresses: rafael.mello@ufrpe.br, rafaelflmello@gmail.com (R. Ferreira), gdcc@cin.ufpe.br (G.D.C. Cavalcanti), fred@cin.ufpe.br (F. Freitas), rdl@cin.ufpe.br (R.D. Lins), steven.simske@hp.com (S.J. Simske), marcelo.riss@hp.com (M. Riss).

http://dx.doi.org/10.1016/j.csl.2017.07.002
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

are semantically equivalent or not (Das and Smith, 2009). Automatic text summarization (Ferreira et al., 2013), information retrieval (Yu et al., 2009), image retrieval (Coelho et al., 2004), text categorization (Liu and Guo, 2005), and machine translation (Papineni et al., 2002) are examples of applications that rely on or may benefit from sentence similarity and PI methods.

The literature reports several efforts to address such a problem by extracting syntactic information from sentences (Islam and Inkpen, 2008; Oliva et al., 2011) or by representing sentences using vectors of bag of words (Mihalcea et al., 2006; Qiu et al., 2006). Sentences are modelled in such a way to allow similarity methods to compute different measures to evaluate the degree of similarity between words. In general, a PI method conveys these similarities as input to machine learning algorithms in order to identify paraphrases. However, two important problems are not handled in traditional sentence similarities approaches:

The Meaning Problem (Choudhary and Bhattacharyya, 2002): It is characterized by the lack of semantic analysis in the previous sentence similarity measures proposed. Essentially this problem is to measure the similarity between the meaning of sentences (Choudhary and Bhattacharyya, 2002). Nevertheless, the measures that claim to deal with it only apply methods such as the latent semantic indexing (Deerwester et al., 1990), corpus-based methods (Li et al., 2003) and WordNet similarity measures (Mihalcea et al., 2006). These techniques, however, are used to find the semantic similarity of the words in a sentence, but not the similarity between two complete sentences. Thus, the evaluation of the meaning similarity degree between two sentences remains still an open problem. For example, the sentences *Peter is a handsome boy* and *Peter is a good-looking lad*, share a similar meaning, if the context they appear in does not change much.

The Word Order Problem (Zhou et al., 2010): In many cases different word order implies in divergent sentences' meanings (Zhou et al., 2010). For example, *A loves B* and *B loves A*, represent two completely different sentences in meaning. Therefore, dealing with this problem certainly enhances the final measure of sentence similarity.

This paper proposes a paraphrase identification system that combines lexical, syntactic and semantic similarity measures. Since traditional methods only rely on lexical and syntactic measures, we believe that the addition of semantic role annotation analysis (Màrquez et al., 2008) is a promising alternative to address the meaning and the word order problems. These three measures were previously tried on and obtained good results on the sentence similarity problem (Ferreira et al., 2014b). The same authors improved their sentence similarity measure by using a similarity matrix that penalizes the measure based on sentence size (Ferreira et al., 2014a). This penalization is important because large sentences could be considered similar even if they contain more information than small sentences. To the best of our knowledge, these measures were never used to identify paraphrases. Thus, the main novelty of this paper is the application of different machine learning algorithms to combine sentence similarity measures in order to identify paraphrases. In addition it presents the concept of Basic Unit to the sentence similarity measures proposed in previous papers.

The proposes system is composed of three steps:

1. *Sentence Representation*: This step performs the lexical, syntactic and semantic analysis and encapsules the outputs in a text file (for lexical) and two RDFs[1] files (for syntactic and semantic).
2. *Similarity Analysis*: It measures the similarity of each pair of sentences using the output of the previous step.
3. *Paraphrase Classification*: The last step applies a machine learning algorithm, using the sentences similarities measures from second step, to identify if the pair of sentences is paraphrase or not.

In order to evaluate the proposed system, a series of experiments was performed using the Microsoft Research Paraphrase Corpus (MSRP) (Dolan et al., 2004), which is the standard dataset for this problem. The proposed approach was compared using four measures: accuracy, precision, recall, and F-measure (Achananuparp et al., 2008), in the experimental study, the principal hypothesis of this work was validated showing that the combination of lexical, syntactic, and semantic aspects of a sentence pair achieve better results for the PI task than state-of-the-art methods. In addition, it is also validated that the use of the sentence representation proposed in (Ferreira et al., 2014b) achieves good performance for the PI task.

[1] Resource Description Framework.