Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 39 (2016) 1–28

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Assessing sentence similarity through lexical, syntactic and semantic analysis☆

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

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

Received 28 April 2015; received in revised form 19 January 2016; accepted 20 January 2016
Available online 6 February 2016

Abstract

The degree of similarity between sentences is assessed by sentence similarity methods. Sentence similarity methods play an important role in areas such as summarization, search, and categorization of texts, machine translation, etc. The current methods for assessing sentence similarity are based only on the similarity between the words in the sentences. Such methods either represent sentences as bag of words vectors or are restricted to the syntactic information of the sentences. Two important problems in language understanding are not addressed by such strategies: the word order and the meaning of the sentence as a whole. The new sentence similarity assessment measure presented here largely improves and refines a recently published method that takes into account the lexical, syntactic and semantic components of sentences. The new method was benchmarked using Li–McLean, showing that it outperforms the state of the art systems and achieves results comparable to the evaluation made by humans. Besides that, the method proposed was extensively tested using the SemEval 2012 sentence similarity test set and in the evaluation of the degree of similarity between summaries using the CNN-corpus. In both cases, the measure proposed here was proved effective and useful.

Keywords: Graph-based model; Sentence simplification; Relation extraction; Inductive logic programming

1. Introduction

The degree of similarity between sentences is measured by sentence similarity or short-text similarity methods. Sentence similarity is important in a number of different tasks, such as: 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). Sentence similarity methods should also be capable of measuring the degree of likeliness between sentences with partial information, as when one sentence is split into two or more short texts and phrases that contain two or more sentences.

☆ This paper has been recommended for acceptance by Srinivas Bangalore.

* Corresponding author. Tel.: +55 81999008818.
E-mail addresses: rflm@cin.ufpe.br (R. Ferreira), rdl@cin.ufpe.br (R.D. Lins), steven.simske@hp.com (S.J. Simske), fred@cin.ufpe.br (F. Freitas), marcelo.riss@hp.com (M. Riss).

http://dx.doi.org/10.1016/j.csl.2016.01.003
0885-2308/

The technical literature reports several efforts to address such problem by representing sentences using a bag of words vector (Mihalcea et al., 2006; Qiu et al., 2006) or a tree of the syntactic information among words (Islam and Inkpen, 2008; Oliva et al., 2011). These representations allow the similarity methods to compute different measures to evaluate the degree of similarity between words. The overall similarity of the sentence is obtained as a function of those partial measures. Two important problems are not handled by using such approach:

The Meaning Problem (Choudhary and Bhattacharyya, 2002) Sentences with the same meaning, but built with different words. For example, the sentences *Peter is a handsome boy* and *Peter is a good-looking lad*, have similar meaning, if the context they appear in does not change much.
The Word Order Problem (Zhou et al., 2010) The order that the words appear in the text influences the meaning of texts. For example, in the sentences "*A killed B*" and "*B killed A*" use the same words, but the order they appear changes their meaning completely.

A recent paper (Ferreira et al., 2014) addressed these problems by proposing a sentence representation and content similarity measure based on lexical, syntactic and semantic analysis. It has some limitations, however. For example, the size of sentences is not taken into account. To overcome such problems, the paper (Ferreira et al., 2014) presents:

- A new sentence representation that improves the one proposed in Ref. Ferreira et al. (2014) to deal with the meaning and word order problems, and
- A sentence similarity measure based on two similarity matrices and a size penalization coefficient.
- An algorithm to combine the statistical and semantic word similarity measures.

This paper, besides explaining the measure presented in Ref. Ferreira et al. (2014) in full details, improves the combination of the word similarity measures, introducing the more general concept of sentence similarity as a numerical matrix. Here, the lexical analysis is performed in the first layer, in which the similarity measure uses "bag-of-word vectors", similarly to Refs. Islam and Inkpen (2008), Li et al. (2006), Mihalcea et al. (2006), Oliva et al. (2011). In addition to lexical analysis, this layer applies two preprocessing services (Hotho et al., 2005): *stopwords removal* and *stemming*. The syntactic layer uses relations to represent the word order problem. The semantic layer employs Semantic Role Annotation (SRA) (Das et al., 2010) to handle both problems. The SRA analysis returns the meaning of the actions, the agent/actor who performs the action, and the object/actor that suffers the action, among other information. Ref. Ferreira et al. (2014) was possibly the first to use SRA as a measure of the semantic similarity between sentences, while other methods employ only WordNet (Fellbaum, 1998; Das and Smith, 2009; Oliva et al., 2011) or a corpus-based measure (Mihalcea et al., 2006; Islam and Inkpen, 2008) in the classic bag-of-word vectors approach.

The measure presented here was benchmarked using three datasets. The one proposed by Li et al. (2006) is widely acknowledged as the standard dataset for such problems. Pearson's correlation coefficient (r) and Spearman's rank correlation coefficient (ρ), which are traditional measures for assessing sentence similarity, were compared with the best results of the measure described in the literature. The new measure proposed in this paper outperforms all the previous ones, as a combination of the proposed measure achieves 0.92 for the r, which means that the proposed measure has the same accuracy of the best human assigned values to the similarities in such a dataset. Compared with ρ, the measure proposed here achieved 0.94, which means a reduction of 33% in the error rate in relation to the state of the art results reported in Ref. Ferreira et al. (2014).

The second experiment described here uses the test set of the SemEval 2012 competition (Agirre et al., 2012), which contains 3108 pairs of sentences. The evaluation was performed in terms of r, which is the official measure used in the competition. The proposed approach obtained 0.6548 for r, only 0.0225 less than the best result reported. However, the approach presented here uses an unsupervised algorithm; the other better ranked systems use supervised algorithms, and are therefore corpus dependent.

The benchmarking experiments also used an extension of the extractive summary datasets in the CNN-corpus proposed by Lins et al. (2012). This corpus is based on CNN news articles from all over the world. The current version of the CNN dataset has 1330 texts in English. One outstanding point of the CNN-corpus is that there is a summary of each text provided by the original author: the *highlights*. The assessment of summary similarity checks the degree of similarity of each sentence in the original text with each of the sentences in the *highlights*. The sentences with the highest similarity scores are seen as providing an extractive summary of the text. Such summary and the highlights are