Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 40 (2016) 8–17

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

CoTO: A novel approach for fuzzy aggregation of semantic similarity measures

Action editor: Péter Érdi

Jorge Martinez-Gil [*]

Software Competence Center Hagenberg, Austria

Received 5 February 2015; received in revised form 25 November 2015; accepted 31 January 2016
Available online 18 February 2016

Abstract

Semantic similarity measurement aims to determine the likeness between two text expressions that use different lexicographies for representing the same real object or idea. There are a lot of semantic similarity measures for addressing this problem. However, the best results have been achieved when aggregating a number of simple similarity measures. This means that after the various similarity values have been calculated, the overall similarity for a pair of text expressions is computed using an aggregation function of these individual semantic similarity values. This aggregation is often computed by means of statistical functions. In this work, we present CoTO (Consensus or Trade-Off) a solution based on fuzzy logic that is able to outperform these traditional approaches.

Keywords: Knowledge-based analysis; Text mining; Semantic similarity measurement; Fuzzy logic

1. Introduction

Semantic similarity measurement is a research challenge whereby two terms or text expressions are assigned a score based on the likeness of their meaning (Pirro, 2009). Accurately measurement of semantic similarity is considered of great importance in many computer related fields since this process is very important for a number of particular scenarios. The reason is that textual semantic similarity measures can be used for understanding beyond the literal representation of words and sentences. For example, it is possible to automatically identify that some terms (e.g., Finance) could be matched with similar terms (e.g., Economics, Economic Affairs, Financial Affairs, and so on).

Identifying different expressions of the same concept is a key method in a lot of disciplines. For example, we can refer to (a) data clustering where semantic similarity measures are necessary to detect and group the most similar subjects (Batet, 2011), (b) data matching which consists of finding some data that refer to the same concept across different data sources (Martinez-Gil & Aldana-Montes, 2010), (c) data mining where using appropriate semantic similarity measures can help to facilitate both the processes of text classification and pattern discovery in large texts (Couto, Silva, & Coutinho, 2005), or (d) automatic machine translation where the detection of terms pairs expressed in different languages is of vital importance (Costa-Jussa & Banchs, 2011).

Traditionally, this problem has been addressed from two different points of view: semantic similarity and relational similarity. However, there is a common agreement about the scope of each of them (Batet, Sanchez, & Valls, 2010). Semantic similarity states the taxonomic proximity between terms or text expressions (Pirro, 2009). For example, automobile and car are similar because they represent

[*] Tel.: +43 7236 3343 838; fax: +43 7236 3343 888.
E-mail address: jorge.martinez-gil@scch.at.

http://dx.doi.org/10.1016/j.cogsys.2016.01.001
1389-0417/© 2016 Elsevier B.V. All rights reserved.

the same notion concerning means of transport. On the other hand, the more general notion of relational similarity considers relations between terms (Punuru & Chen, 2012). For example, nurse and hospital are related (since they belong to the healthcare domain) but they are far from represent the same real idea or concept. Due to its importance in many computer-related fields, we are going to focus on semantic similarity for the rest of this paper.

There are many methods for identifying semantic similarity. However, the best results have been often achieved when aggregating a number of simple similarity measures (Do & Rahm, 2002). This means that after the various semantic similarity values have been achieved, the final similarity score for two text expressions is computed using an aggregation function of the individual semantic similarity values. This aggregation process is often computed by means of statistical functions (arithmetic mean, quadratic mean, median, maximum, minimum, and so on) (Martinez-Gil & Aldana-Montes, 2012). We think that these methods are not optimal, and therefore, results can be improved. The reason is that these methods are following a kind of compensative approach, and therefore they are not able to deal with the non-stochastic uncertainty induced from subjectivity, vagueness and imprecision from the humans when using their languages. We think that using a fuzzy operator should help to outperform current results in the field of semantic similarity measurement. Therefore, the key contributions of this work can be summarized as follows:

- We propose CoTO (Consensus or Trade-Off), a fuzzy operator for the aggregation of semantic similarity values that appropriately handles the non-stochastic uncertainty inherent to human language.
- We evaluate the performance of this strategy using a number of general purpose and domain specific benchmark data sets, and show how this new approach outperforms the results from existing techniques.

The rest of this paper is organized as follows: Section 2 describes the state-of-the-art concerning semantic similarity measurement. Section 3 describes the novel approach for the fuzzy aggregation of simple semantic similarity measures. Section 4 describes our evaluations and the results that have been achieved. Finally, we draw conclusions and put forward future lines of research.

2. Related work

Textual semantic similarity represents a widely intuitive concept. Miller and Charles wrote: *…subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty* (Miller & Charles, 1991). This viewpoint has been reinforced by other researchers in the field who observed that semantic similarity is treated as a property characterized by human perception and intuition (Resnik, 1999). In general, it is assumed that not only are the participants comfortable in their understanding of the concept, but also when they perform a judgment task they do it using the same procedure or at least have a common understanding of the attribute they are measuring (O'Shea, Bandar, Crockett, & McLean, 2010).

In the past, there have been great efforts in finding new semantic similarity measures mainly due it is of fundamental importance in many application-oriented fields of the modern computer science. The reason is that these techniques can be used for going beyond the literal lexical match of words and text expressions. Past works in this field include the automatic processing of text messages (Lamontagne & Lapalme, 2004), healthcare dialogue systems (Bickmore & Giorgino, 2006), natural language querying of databases (Erozel, Cicekli, & Cicekli, 2008) and question answering (Moschitti & Quarteroni, 2008).

On the other hand, according to Sanchez, Batet, and Isern (2011); most of these existing semantic similarity measures can be classified into one of these four main categories.

1. Edge-counting measures which are based on the computation of the number of taxonomical links separating two concepts represented in a given dictionary (Leacock & Chodorow, 1998).
2. Feature-based measures which try to estimate the amount of common and non-common taxonomical information retrieved from dictionaries (Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2003).
3. Information theoretic measures which try to determine similarity between concepts as a function of what both concepts have in common in a given ontology. These measures are typically computed from concept distribution in text corpora (Jiang & Conrath, 1997).
4. Distributional measures which use text corpora as source. They look for word co-occurrences in the Web or large document collections using search engines (Bollegala, Matsuo, & Ishizuka, 2011).

It is not possible to categorize CoTO into any of these categories since we are not proposing a new semantic similarity measure, but a novel method to aggregate them so that individual measures can be outperformed. In this way, semantic similarity measures are like black boxes for us. However, there are several related works in the field of semantic similarity aggregation. For instance COMA, where a library of semantic similarity measures and friendly user interface to aggregate them are provided (Do & Rahm, 2002), or MaF, a matching framework that allow users to combine simple similarity measures to create more complex ones (Martinez-Gil & Aldana-Montes, 2011).

These approaches can be even improved by using weighted means where the weights are automatically computed by means of heuristic and meta-heuristic algorithms.