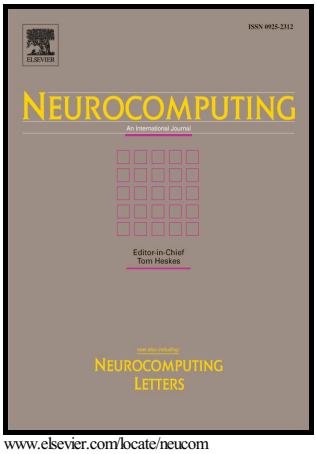

www.elsevier.com/locate/neucom

LWCR: Multi-Layered Wikipedia representation for Computing word Relatedness

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, Abdelmajid Ben Hamadou

PII: S0925-2312(16)30941-9
DOI: http://dx.doi.org/10.1016/j.neucom.2016.08.045
Reference: NEUCOM17479

To appear in: *Neurocomputing*

Received date: 14 July 2015
Revised date: 26 February 2016
Accepted date: 15 August 2016

Cite this article as: Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb and Abdelmajid Ben Hamadou, LWCR: Multi-Layered Wikipedia representation for Computing word Relatedness, *Neurocomputing* http://dx.doi.org/10.1016/j.neucom.2016.08.045

LWCR: multi-Layered Wikipedia representation for Computing word Relatedness

Mohamed Ben Aouicha[a,d], Mohamed Ali Hadj Taieb[b,d,*], Abdelmajid Ben Hamadou[c,d]

[a]Faculty of Sciences, Sfax University, Tunisia
[b]Higher Institute of Applied Sciences and Technology, Sousse University, Tunisia
[c]Higher Institute of Computer Science and Multimedia, Sfax University, Tunisia
[d]Multimedia InfoRmation system and Advanced Computing Laboratory, Sfax University, Sfax 3021, Tunisia
[*] Corresponding author Tel.: +216 24688354. E-mail address: mohamedali.hadjtaieb@gmail.com

Abstract

The measurement of the semantic relatedness between words has gained increasing interest in several research fields, including cognitive science, artificial intelligence, biology, and linguistics. The development of efficient measures is based on knowledge resources, such as Wikipedia, a huge and living encyclopedia supplied by net surfers. In this paper, we propose a novel approach based on multi-Layered Wikipedia representation for Computing word Relatedness (LWCR) exploiting a weighting scheme based on Wikipedia Category Graph (WCG): Term Frequency-Inverse Category Frequency (tfxicf). Our proposal provides for each category pertaining to the WCG a Category Description Vector (CDV) including the weights of stems extracted from articles assigned to a category. The semantic relatedness degree is computed using the cosine measure between the CDVs assigned to the target words couple. The basic idea is followed by enhancement modules exploiting other Wikipedia features, such as article titles, redirection mechanism, and neighborhood category enrichment, to exploit semantic features and better quantify the semantic relatedness between words. To the best of our knowledge, this is the first attempt to incorporate the WCG-based term-weighting scheme (tfxicf) into computing model of semantic relatedness. It is also the first work that exploits 17 datasets in the assessment process, which are divided into two sets. The first set includes the ones designed for semantic similarity purposes: RG65, MC30, AG203, WP300, SimLexNoun666 and GeReSiD50Sim; the second includes datasets for semantic relatedness evaluation: WordSim353, GM30, Zeigler25, Zeigler30, MTurk287, MTurk771, MEN3000, Rel122, ReWord26, GeReSiD50 and SCWS1229. The found results are compared to WordNet-based measures and distributional measures cosine and PMI performed on Wikipedia articles. Experiments show that our approach provides consistent improvements over the state of the art results on multiple benchmarks.

Keywords

Semantic analysis; Wikipedia category graph; tfxicf; co-occurrence; vector-based measure, semantic relatedness

I. Introduction

Expressing the meaning of a word passes mainly by two axes: the first concerns the hierarchical structure that organizes the words in meaningful way and the second is the content where it cohabits with other words forming the target sense. The sense quantification is a key element for computing the