Knowledge-Based Systems 50 (2013) 260–278

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Computing semantic relatedness using Wikipedia features

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

Multimedia InfoRmation System and Advanced Computing Laboratory, Sfax University, Sfax 3021, Tunisia

ARTICLE INFO

Article history:
Received 23 January 2013
Received in revised form 14 June 2013
Accepted 20 June 2013
Available online 4 July 2013

Keywords:
Semantic relatedness
Wikipedia
Wikipedia category graph
Word relatedness
Semantic analysis

ABSTRACT

Measuring semantic relatedness is a critical task in many domains such as psychology, biology, linguistics, cognitive science and artificial intelligence. In this paper, we propose a novel system for computing semantic relatedness between words. Recent approaches have exploited Wikipedia as a huge semantic resource that showed good performances. Therefore, we utilized the Wikipedia features (articles, categories, Wikipedia category graph and redirection) in a system combining this Wikipedia semantic information in its different components. The approach is preceded by a pre-processing step to provide for each category pertaining to the Wikipedia category graph a semantic description vector including the weights of stems extracted from articles assigned to the target category. Next, for each candidate word, we collect its categories set using an algorithm for categories extraction from the Wikipedia category graph. Then, we compute the semantic relatedness degree using existing vector similarity metrics (Dice, Overlap and Cosine) and a new proposed metric that performed well as cosine formula. The basic system is followed by a set of modules in order to exploit Wikipedia features to quantify better as possible the semantic relatedness between words. We evaluate our measure based on two tasks: comparison with human judgments using five datasets and a specific application "solving choice problem". Our result system shows a good performance and outperforms sometimes ESA (Explicit Semantic Analysis) and TSA (Temporal Semantic Analysis) approaches.

1. Introduction

Semantic **R**elatedness **(SR)** is used as a necessary pre-processing step to many Natural Language Processing (NLP) tasks, such as Word Sense Disambiguation (WSD) [21,15]. Moreover, SR constitutes one of the major stakes in the Information Retrieval (IR) [10,2,13,56,60] especially in some tasks such as semantic indexing [51]. A powerful semantic relatedness measure can have influences on Semantic Information Retrieval (SIR) system. It actually exists in some information retrieval systems that support retrieval by Semantic Similarity Retrieval Model (SSRM) [47]. Research in semantic technologies has had a major impact by enabling and improving a wide range of web-based applications, such as search [8] as well as category discovery of videos [44].

Semantic relatedness measures typically use linguistic knowledge resources like WordNet[1] [9] whose construction is very expensive and time-consuming. So far, insufficient coverage of these linguistic resources has been a major impediment for using semantic relatedness measures in large-scale natural language processing applications. Some of these rapidly growing collaboratively constructed resources like Wikipedia have the potential to be used as a new kind of semantic resource due to their increasing size and significant coverage. Wikipedia has recently been widely recognized as an enabling knowledge base for a variety of intelligent systems would be useful to define semantic relatedness [11,25,29,35,43,50].

This work exploits Wikipedia features for measuring semantic relatedness between words. The rest of the paper is organized as follows. Section 2 gives a detailed overview of the state of the art in computing semantic relatedness using Wikipedia except the temporal semantic analysis which uses the archive of The New York Times spanning 150 years. Section 3 describes the semantic relatedness system that must be preceded by a pre-processing step to generate the category semantic depiction used to extract categories assigned to the candidate couple words. Section 4 presents the evaluating semantic relatedness measures and different benchmarks formed by human judgments. Section 5 details our system modules exploiting different Wikipedia features. In this section the theoretical aspect is presented in parallel with the experimental side to show the contribution of each enhancement. As for Section 6, it is a synthesis section to summarize the performance of the result system on the different datasets. Moreover, we discuss within the same section our system in comparison with the existing SR measures. Section 7 includes the evaluation of the effectiveness of our proposed method in solving the word choice

* Corresponding author. Tel.: +216 24 688 354.
E-mail addresses: mohamedali.hadjtaieb@gmail.com (M.A. Hadj Taieb), benaouicha.mohamed@gmail.com (M. Ben Aouicha), abdelmajid.benhamadou@isimsf.rnu.tn (A. Ben Hamadou).

[1] http://wordnet.princeton.edu/

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.06.015

problem task. Concluding remarks and some future directions of our work are described in Section 8.

2. Related works

Several researches have been done to use Wikipedia as a semantic resource for computing the semantic relatedness between words or concepts. In this section, we present some main approaches.

2.1. Wikirelate!

Wikirelate! ceated by Strube and Ponzetto [43] is based on category structure. Their system works as follows: given the word pairs (w_i, w_j); they first retrieve the Wikipedia pages which they refer to. Then, they run through the category tree to extract the categories that the pages belong to. Finally, they compute relatedness based on the pages extracted and the paths connecting categories in the category taxonomy.

2.1.1. Page retrieval and disambiguation

Page retrieval for page p_i is accomplished by first querying the page titled as the w_i. Next, they follow all redirects (i.e. CAR redirecting to AUTOMOBILE) and resolve ambiguous page queries, as many queries to Wikipedia return a disambiguation page which contains many hyperlinks as candidate targets for the given original query.

To disambiguate the page p_i for w_i, they first get all the hyperlinks in page p_j obtained for w_j without disambiguating. This is to bootstrap the disambiguation process. It could also be the case that both queries are ambiguous. They take the other w_j and all the Wikipedia internal links of the page p_j as a lexical association list to be used for disambiguation. If a link in p_i contains any occurrence of a disambiguating term, the target page is returned; else we return the first article linked in the disambiguation page. This disambiguation strategy offers a less accurate solution than following all disambiguation page links. Nevertheless it offers a more practical solution as many of those pages contain a large number of links.

2.1.2. Category tree search

Paths along the category tree are needed for computing path and information content based measures. Given the pages p_i and p_j, they extract the lists of categories C_i and C_j to which they belong. The category links in the pages are considered as primitive concepts in the taxonomy which the words denote. Given the category lists, for each category pair they perform a depth-limited search of maximum depth of 4 for a least common subsumer. They approve that limiting the search improves the results. This is due to the strongly connected Wikipedia Category Graph (WCG).

2.1.3. Relatedness measure computation

Finally, given the set of paths found between the category pairs, they use the taxonomy based measures, namely path based measures and information content based measures.

2.2. Explicit semantic analysis

Explicit Semantic Analysis (ESA as illustrated in Fig. 1) created by Gabrilovich and Markovitch [11] is based on text features and link within articles. They use this latter association-based method to assign semantic interpretation to words and text fragments. They assume the availability of a vector of basic concepts, $C_1, \ldots, C_n$, and they represent each text fragment t by a vector of weights, $w_1, \ldots, w_n$, where w_i represents the strength of association between t and C_i. This weighted vector is called the semantic interpretation vector of t.

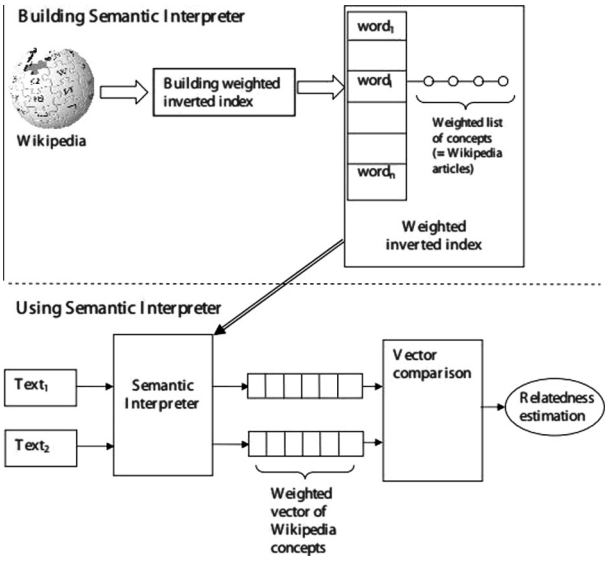

Fig. 1. The ESA system structure.

Gabrilovich and Markovitch use Wikipedia articles as index documents since Wikipedia covers a wide range of topics, while each article is focused on one topic. Each Wikipedia concept is represented as a vector of words that occur in the corresponding article. Entries of these vectors are assigned weights using *tf X idf* scheme [39]. These weights quantify the strength of association between words and concepts.

They build an inverted index, which maps each word into a list of concepts in which it appears. They also use the inverted index to discard insignificant associations between words and concepts by removing those concepts whose weights for a given word are too low. To compute semantic relatedness of a pair of words they compare their vectors using the cosine metric.

Meanwhile, ESA has been adopted successfully in many applications. In [7], Egozi et al. show that ESA contributes directly to the estimation of the relevance of documents for a given query. In other settings ESA is used to compute the semantic relatedness of terms [26], which are then used as parameters in other retrieval models (e.g., an extension of BM-25). A cross-lingual extension (CL-ESA) that exploits interlanguage links of Wikipedia articles is covered in [33] and [41]. In [4], Cimiano et al. show that CL-ESA is superior to other retrieval models which are based on implicit semantics.

2.3. Wikipedia Link Vector Model

Wikipedia Link Vector Model (WLVM) approach created by Milne [25] extracts the semantic relatedness measures for term pairs from the Wikipedia's link structure. The first step concerns the extraction of Wikipedia articles in relation with the word pair (w_1, w_2). The best method to obtain such articles is to look for the Wikipedia pages that directly match the term. Thus, to avoid problems caused by the ambiguous terms, the procedure to obtain the related articles is listing all the pages which titles match the term. In fact, they take all articles that contain the term in their titles with the corresponding articles of redirects and target articles existing in disambiguation pages.

The next step is to judge the semantic similarity between the articles obtained in the first step. This similarity between the two Wikipedia articles can be defined as the angle between the vectors of the links found within them. Thus, if k is the total number of articles within Wikipedia then the weighted value w for the link $a \rightarrow b$ is: