YJBIN 2483
19 December 2015
ARTICLE IN PRESS
No. of Pages 18, Model 5G

Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Computing semantic similarity between biomedical concepts using new information content approach

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb *

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

ARTICLE INFO

Article history:
Received 11 June 2015
Revised 6 November 2015
Accepted 12 December 2015
Available online xxxx

Keywords:
Semantic similarity
Information content
MeSH
Biomedicine

ABSTRACT

The exploitation of heterogeneous clinical sources and healthcare records is fundamental in clinical and translational research. The determination of semantic similarity between word pairs is an important component of text understanding that enables the processing and structuring of textual resources. Some of these measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data or from medical ontologies such as MeSH. This study focuses on Information Content (IC) based measures that exploit the topological parameters of the taxonomy to express the semantics of a concept. A new intrinsic IC computing method based on the taxonomical parameters of the ancestors' subgraph is then assigned to a biomedical concept into the "*is a*" hierarchy. Moreover, we present a study of the topological parameters through the MeSH taxonomy. This study treats the semantic interpretation and the different ways of expressing the parameters of depth and the descendants' subgraph. Using MeSH as an input ontology, the accuracy of our proposal is evaluated and compared against other IC-based measures according to several widely-used benchmarks of biomedical terms. The correlation between the results obtained for the evaluated measure using the proposed approach and those from the ratings of human' experts shows that our proposal outperforms the previous measures.

1. Introduction

Clinicians are confronted with increasing amounts of medical data from multiple sources housed in electronic format. The huge amounts of clinical and scientific documents in digital libraries and the digitized records assigned to patient health are valuable resources for clinical and translational research. Translational research includes medical information on patient health that comes from various sources and systems, including empirical observations, visits, and worksheet. The provided information is often heterogeneous and unprocessed. There is increasing interest in recent research in the search for variable strategies to manage and process this huge flow of data. The literature indicates that semantic technology offers promising opportunities for the development of efficient approaches to the interpretation of data from multiple origins and for determining the relationship between them.

The estimation of the semantic similarity between words is one of the major tools employed in semantic technology for text processing and understanding. It has been widely applied in several natural language processing tasks, such as word sense disambiguation [1,2], document categorization or clustering [3,4], word spelling correction [5], automatic language translation [4], ontology learning [6], and information retrieval [7,8].

In the biomedical field, the computation of the similarity between words can improve the performance of information retrieval from biomedical sources [8,9], integration of heterogeneous clinical data [10], automation of semantic grouping of clinical word pairs [11], and clustering of clinical models from local electronic health records [12].

Semantic similarity is a computational method used to identify and quantify likeness between words using the common characteristics shared between them. For example, *bronchitis* and *flu* are similar because they are both disorders of the respiratory system. The semantic similarity is based on the evaluation of the semantic evidence observed in a knowledge source (such as ontologies or domain corpora). According to the type of domain knowledge exploited, different families of functions can be identified: those based on the taxonomical structure of an ontology and those relying on the intrinsic Information Content (IC) of concepts [13–18].

* Corresponding author.
E-mail address: mohamedali.hadjtaieb@gmail.com (M.A.H. Taieb).

http://dx.doi.org/10.1016/j.jbi.2015.12.007
1532-0464/© 2015 Published by Elsevier Inc.

Please cite this article in press as: M.B. Aouicha, M.A.H. Taieb, Computing semantic similarity between biomedical concepts using new information content approach, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.12.007

These measures perform poorly with biomedical terms if they are exploited with general purpose knowledge [19], such as WordNet[1] [20]. The problem with WordNet is not the quality of the semantic relations between the present biomedical concepts, but with its coverage capacity (only 25.1% of MeSH terms are covered in WordNet [19]). Therefore, there are a number of relevant biomedical ontologies, knowledge repositories and structured vocabularies that model and organize concepts in a comprehensive way. Well-known examples are MeSH (Medical Subject Headings) for indexing literature, the ICD taxonomy (International Classification of Diseases) for recording causes of death and diseases, and SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) the most comprehensive and precise clinical health terminology product, owned and distributed around the world by The International Health Terminology Standards Development Organisation (IHTSDO). Several similarity computation approaches have been compared using the biomedical knowledge source and evaluated over particular datasets or in the context of a concrete application, such as document clustering [3,21], assessment of similarity between words [8,22–24], and providing a useful basis for assessing the structure of terminological systems and the content of medical records [25].

In this paper, we first review and discuss the IC-based semantic similarity measures commonly referenced in the literature and provide details of their potential adaptation to the biomedical domain. We also analyze the taxonomic parameters used for the quantification of intrinsic information content of biomedical concepts to determine their semantic interpretations. In order to overcome some of the problems identified in this study, we present a new intrinsic IC computing method based on the exploitation of the ancestors' subgraph and the quantification of the specificity of each hypernym. Finally, the paper evaluates and compares the results obtained by our measure against those reported by other similarity functions when applied to the biomedical domain. The results show that our proposed method, coupled with Lin's similarity measure, displays a high level of correlation and outperforms other IC computing approaches.

The rest of the paper is organized as follows. Section 2 presents a survey about the IC-based semantic similarity measures, including the IC computing methods and the similarity measures. Section 3 provides a study of the topologic parameters extracted from the MeSH taxonomic knowledge resource for the computation of semantic similarity between biomedical concepts. Section 4 describes the new intrinsic IC-computing method of a biomedical concept based on its ancestors' subgraph and the taxonomic parameters. Section 5 reports on the evaluation and comparison of our approach against currently available ones using known benchmarks and the biomedical resource MeSH. The final section is devoted to presenting our conclusions and recommendations for future research.

2. Related works: information content-based semantic similarity measures

The measurement of semantic similarity based on Information Content (IC) was first introduced by Resnik [1]. The basic idea of IC is that general and abstract entities found in a discourse present less IC than more concrete and specialized ones. This principle is inspired from the work of Shannon [26]. The more probable a concept appears, the less information it conveys. In other words, specific words are more informative than general ones. IC-based semantic similarity measures [27–29] consist of two parts: the **computing IC method** and the **IC-based measure**. There are two ways for quantifying IC: the first exploits corpora, and the second, which is often described as *intrinsic*, uses topological parameters from the hierarchical knowledge structure: descendants (hyponyms), depth, leaves, and ancestors (hypernyms), for quantifying the IC of a concept. The terms "*hypernym/hyponym*", "*ancestors/descendants*" and "*subsumers*" are used as follows:

– *Hypernym/hyponym:* In the "*is a*" relation linking two concepts, such as "*Animal*" and "*Pet*", "*Animal*" is called hypernym of "*Pet*", and "*Pet*" is an hyponym of "*Animal*".
– *Ancestors/descendants:* ancestors of a concept pertaining to "*is a*" hierarchy refer to direct and indirect hypernyms. Descendants refer to direct and indirect hyponyms.
– Subsumer: a concept c_1 is a subsumer of c_2 if c_2 is a descendant of c_1.

IC-based similarity measures exploit the IC-values assigned to concepts c_1 and c_2 to provide the semantic similarity estimation between them. A complete survey of IC-based similarity measures is presented in the next paragraph.

2.1. Similarity measures exploiting the IC

Several semantic similarity measures, which are based on the exploitation of the information content, have been proposed. The similarity estimation between two concepts c_1 and c_2 is computed using their ICs and the IC of the Lowest Common Subsumer (LCS) which is extracted from the "*is a*" hierarchy. Some measures are presented in next paragraphs:

- *Resnik:* Guided by the idea that the similarity between a pair of concepts may be judged by "the amount of shared information", Resnik [1] defined the similarity between two concepts as the IC of their Lowest Common Subsumer $LCS(c_1,c_2)$ as follows:

$$\mathrm{Sim}_{\mathrm{Res}}(c_1,c_2)=\mathrm{IC}(\mathrm{LCS}(c_1,c_2)) \tag{1}$$

- *Jiang-Conrath:* This approach subtracts the IC of the LCS from the sum of the IC of the individual concepts [30]. It provides the dissimilarity estimation between two terms, because the more different the terms are, the higher the difference between their ICs and the IC of their LCS will be. The dissimilarity measure is expressed as follows:

$$\mathrm{Dis}_{\mathrm{JC}}(c_1,c_2)=(\mathrm{IC}(c_1)+\mathrm{IC}(c_2))-2\mathrm{IC}(\mathrm{LCS}(c_1,c_2)) \tag{2}$$

- *Lin:* The similarity measure described by Lin [31] is defined as Dice coefficient:

$$\mathrm{Sim}_{\mathrm{Lin}}(c_1,c_2)=\frac{2\times\mathrm{IC}(\mathrm{LCS}(c_1,c_2))}{\mathrm{IC}(c_1)+\mathrm{IC}(c_2)} \tag{3}$$

- *Pirro:* He proposes a similarity measure [23] that is conceptually similar to the JC and Lin measures. However, it is based on the feature-based theory of similarity described by Tversky [32]. According to Tversky, the similarity between two concepts c_1 and c_2 is a function of the features common to c_1 and c_2, those in c_1 but not in c_2, and those in c_2 but not in c_1. The semantic similarity between concepts can be computed as an aggregation between the ICs of c_1, c_2, and their LCS:

$$\mathrm{Sim}_{\mathrm{tvr}}(c_1,c_2)=3\times\mathrm{IC}(\mathrm{LCS}(c_1,c_2))-\mathrm{IC}(c_1)-\mathrm{IC}(c_2) \tag{4}$$

Finally, the measure is defined as follows:

$$\mathrm{Sim}_{\mathrm{P\&S}}(c_1,c_2)=\begin{cases}\mathrm{Sim}_{\mathrm{tvr}}(c_1,c_2) & \text{if } c_1\neq c_2\\ 1 & \text{if } c_1=c_2\end{cases} \tag{5}$$

- *Meng:* This measure [33] used Lin's measure. It increases monotonically with $\mathrm{Sim}_{\mathrm{Lin}}$ as follows:

$$\mathrm{Sim}_{\mathrm{Meng}}(c_1,c_2)=e^{\mathrm{Sim}_{\mathrm{Lin}}(c_1,c_2)}-1 \tag{6}$$

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