Journal of Biomedical Informatics 54 (2015) 329–336

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs

Bridget T. McInnes [a,*], Ted Pedersen [b]

[a] Department of Computer Science, Virginia Commonwealth University, 401 S. Main St., Rm E4225, Richmond, VA 23284, USA
[b] Department of Computer Science, University of Minnesota, 1114 Kirby Drive, Duluth, MN 55812, USA

ARTICLE INFO

Article history:
Received 4 June 2014
Accepted 13 November 2014
Available online 15 December 2014

Keywords:
Natural language processing
NLP
Semantic similarity
Semantic relatedness

ABSTRACT

Introduction: This article explores how measures of semantic similarity and relatedness are impacted by the semantic groups to which the concepts they are measuring belong. Our goal is to determine if there are distinctions between homogeneous comparisons (where both concepts belong to the same group) and heterogeneous ones (where the concepts are in different groups). Our hypothesis is that the similarity measures will be significantly affected since they rely on hierarchical *is-a* relations, whereas relatedness measures should be less impacted since they utilize a wider range of relations. In addition, we also evaluate the effect of combining different measures of similarity and relatedness. Our hypothesis is that these combined measures will more closely correlate with human judgment, since they better reflect the rich variety of information humans use when assessing similarity and relatedness.

Method: We evaluate our method on four reference standards. Three of the reference standards were annotated by human judges for relatedness and one was annotated for similarity.

Results: We found significant differences in the correlation of semantic similarity and relatedness measures with human judgment, depending on which semantic groups were involved. We also found that combining a definition based relatedness measure with an information content similarity measure resulted in significant improvements in correlation over individual measures.

Availability: The semantic similarity and relatedness package is an open source program available from http://umls-similarity.sourceforge.net/. The reference standards are available at http://www.people.vcu.edu/~{}btmcinnes/downloads.html.

1. Introduction

Semantic similarity and relatedness measures quantify the degree to which two concepts are similar (e.g., *liver-organ*) or related (e.g., *headache- aspirin*). Relatedness encompasses many kinds of relations, but generally shows how associated two concepts are with each other. For example, a headache can be *treated* with aspirin. Similarity is a specific relation that is a subset of relatedness, and is based on the degree to which two concepts are connected through hierarchical *is-a* relations. For example, *organ* could be an ancestor of *liver* in an *is-a* hierarchy, and would therefore have a high similarity score. *Headache* and *aspirin*, on the other hand, are not closely connected by any *is-a* relations, and so would have a low similarity score. However, since they may be connected by other kinds of relations (e.g., *treated by*) they could have a very high relatedness score.

The automated discovery of groups of semantically similar or related concepts and terms is critical to improving the retrieval [1] and clustering [2] of biomedical and clinical documents, and the development of biomedical terminologies and ontologies [3]. As such, a number of different similarity measures have been developed for the biomedical domain. These have been evaluated intrinsically via comparisons to various human reference standards [4,5], as well as extrinsically depending on how well they contribute to the performance of secondary applications [6,7]. However, to date there has been little work that considers the type of concept being evaluated. Our objective is to evaluate how measures of similarity and relatedness perform depending on the semantic groups of the concepts involved.

Similarity measures find paths between concepts in an *is-a* hierarchy. Concept pairs from different semantic groups may well be in different hierarchies and therefore not be connected by *is-a* relations. In addition, these different hierarchies may have different levels of granularity and coverage. Given these considerations, our hypothesis is that there will be a large degree of change in the correlation of similarity measures with human reference

* Corresponding author.
E-mail addresses: btmcinnes@vcu.edu (B.T. McInnes), tpederse@d.umn.edu (T. Pedersen).

http://dx.doi.org/10.1016/j.jbi.2014.11.014
1532-0464/

standards when the concepts in a pair are from different semantic groups. Our results support this hypothesis. We found that no single measure performed best over all the different semantic group pairs.

In this work, we also combined measures based on the hypothesis that measures of similarity and relatedness will be complementary, and may result in more robust measures that more closely correlate with human judgments. Our goal is to identify pairs of measures that provide complementary information that will improve our ability to quantify the degree of similarity and relatedness between two terms. Bill et al. [8] showed that a linear combination of the similarity measures proposed by Resnik [9] and Lin [10] increased the accuracy of identifying similar terms. The results, here in this paper, show that combining relatedness and similarity measures improved correlation scores overall. However, these results varied depending on the reference standard used and so no single pair of measures was found to always improve correlation.

This article is organized as follows. Section 2 provides an overview of the Unified Medical Language System (UMLS), which is our main source of data on concepts and their relations. Section 3 reviews the measures of semantic similarity and relatedness used in this study. Section 4 describes resources used beyond the UMLS for formulating some of the measures. The reference standards used in our evaluation are introduced in Section 5, and the details of our experiments on these standards are summarized in Section 6. Our results are presented in Section 7, and the article closes with our conclusions in Section 8.

2. Unified Medical Language System

The UMLS is a data warehouse containing three knowledge sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus contains approximately 2 million biomedical and clinical concepts from over 100 different terminologies that have been semi-automatically integrated into a single source. One such source is the *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT), which is a comprehensive clinical terminology created for the electronic representation of clinical health information. The concepts in SNOMED CT are organized in a hierarchical structure in order to permit searching at various levels of specificity. The concepts are connected by two main types of hierarchical relations: *parent/child* (PAR/CHD) and *broader/narrower* (RB/RN). The PAR/CHD relations are strictly *is-a* relations while the RB/RN relations contain *part-of* relations.

The Semantic Network consists of a set of broad subject categories called semantic types in which each concept in the Metathesaurus is assigned one or more semantic type. For example, the semantic type of C0206250 [Autonomic nerve] is *Body Part, Organ, or Organ Component*. Currently, there exist 135 semantic types in the Semantic Network.

The SPECIALIST Lexicon contains terms that are used in the biomedical and health-related domain along with linguistic information such as spelling variants.

Included in the UMLS is also a categorization of semantic types referred to as *semantic groups*. A semantic group is a coarse grained grouping of the semantic types in the UMLS developed by [11] to provide a coarse-grained distinction between UMLS concepts based on their semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. Examples of semantic groups include: Anatomy, Phenomena, Disorders and Chemicals & Drugs. There currently exists 15 semantic groups.[1] Each CUI in the UMLS can be categorized by their semantic group.

[1] http://semanticnetwork.nlm.nih.gov/SemGroups/.

3. Similarity and relatedness measures

This section describes the similarity and relatedness measures used in this work.

3.1. Similarity measures

We classify the similarity measures into two broad categories: path-based and information content (IC)-based. The path-based similarity measures provide information about the co-location of the terms in a taxonomy. The IC measures use the taxonomy information but also include additional information about the concept with respect to its relationship with the other concepts. There are two methods used to calculate IC: *corpus-based* which uses the probability of the concept occurring in an external corpus, and *intrinsic-based* which uses the informativeness of a concept based on its placement within the taxonomy. The remainder of this subsection describes the various measures and how they are calculated.

3.1.1. Path-based measures

Rada et al. [1] introduce the Conceptual Distance measure, which is the length of the shortest path between two concepts ($c1$ and $c2$) in MeSH using RB/RN relations. Caviedes and Cimino [12] later evaluated this measure using the PAR/CHD relations. The *path* measure is a modification of this and is calculated as the reciprocal of the length of the shortest path as defined in Eq. (1).

$$\text{sim}_{path} = \frac{1}{\text{spath}(c_1, c_2)} \tag{1}$$

Wu and Palmer [13] extend this measure by incorporating the depth of the Least Common Subsumer (LCS). The LCS is the most specific ancestor two concepts share. In this measure, the similarity is twice the depth of the two concepts' LCS divided by the product of the depths of the individual concepts as defined in Eq. (2).

$$\text{sim}_{wup} = \frac{2 * \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \tag{2}$$

Leacock and Chodorow [14] extend the path measure by incorporating the depth of the taxonomy. Here, the similarity is the negative log of the shortest path (*spath*) between two concepts divided by twice the total depth of the taxonomy (D) as defined in Eq. (3).

$$\text{sim}_{lch} = -\log \frac{\text{spath}(c_1, c_2)}{2 * D} \tag{3}$$

3.1.2. Information Content (IC) measures

Information content (IC) is formally defined as the negative log of the probability of a concept. Resnik [9] modified IC to be used as a similarity measure. He defined the similarity of two concepts to be the IC of their LCS as shown in Eq. (4).

$$\text{sim}_{res} = \text{IC}(\text{lcs}(c_1, c_2)) = -\log(P(\text{lcs}(c_1, c_2))) \tag{4}$$

Jiang and Conrath [15] and Lin [10] extended Resnik's IC measure by incorporating the IC of the individual concepts. Lin defined the similarity between two concepts by taking the quotient between twice the IC of the concepts' LCS and the sum of the IC of the two concepts as shown in Eq. (5). This is similar to the measure proposed by Wu & Palmer; differing in the use of IC rather than the depth of the concepts.

$$\text{sim}_{lin} = \frac{2 * \text{IC}(\text{lcs}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)} \tag{5}$$

Jiang and Conrath defined the distance between two concepts to be the sum of the IC of the two concepts minus twice the IC of the concepts' LCS. We modify this measure to return a similarity score by taking the reciprocal of the distance as shown in Eq. (6).