Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts

Rimma Pivovarov, Noémie Elhadad *

Department of Biomedical Informatics, Columbia University, 622 W. 168th Street, VC-5, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 13 September 2011
Accepted 8 January 2012
Available online 25 January 2012

Keywords:
Semantic similarity
SNOMED-CT
Distributional semantics
Graph-based metrics
Ontologies

ABSTRACT

An open research question when leveraging ontological knowledge is when to treat different concepts separately from each other and when to aggregate them. For instance, concepts for the terms "paroxysmal cough" and "nocturnal cough" might be aggregated in a kidney disease study, but should be left separate in a pneumonia study. Determining whether two concepts are similar enough to be aggregated can help build better datasets for data mining purposes and avoid signal dilution. Quantifying the similarity among concepts is a difficult task, however, in part because such similarity is context-dependent. We propose a comprehensive method, which computes a similarity score for a concept pair by combining data-driven and ontology-driven knowledge. We demonstrate our method on concepts from SNOMED-CT and on a corpus of clinical notes of patients with chronic kidney disease. By combining information from usage patterns in clinical notes and from ontological structure, the method can prune out concepts that are simply related from those which are semantically similar. When evaluated against a list of concept pairs annotated for similarity, our method reaches an AUC (area under the curve) of 92%.

1. Introduction

A standard way of approaching unstructured biomedical texts, such as patient notes written by clinicians, is to map mentions of biomedical terms, like symptoms and disease names, to semantic concepts in structured and standardized nomenclatures. The mapping helps group all lexical variants of the same biomedical concept under a unique semantic representation, thereby abstracting away from stylistic differences. For instance, the terms "heart attack", "myocardial infarction", and "MI" are all mapped to the same concept in the Unified Medical Language System (UMLS) [1], a conglomerate of different biomedical terminologies. However, most biomedical ontologies and terminologies are designed based on a fine-grained organization of semantic concepts. As a result, when mapping term mentions in a text to semantic concepts, all too often semantically similar terms are mapped to different concepts in the ontology. When the concepts are fed to data mining or pattern recognition analyses, this ontological granularity can result in problems of signal dilution [2]. To enrich the sparse datasets and thus enable meaningful analysis, concepts that are semantically similar can be aggregated. The evaluation of whether two concepts are semantically similar enough for aggregation is often highly dependent on the context of the study itself [3]. For example, concepts such as "obese" and "morbidly obese" can be merged when studying Huntington's disease, but should remain separate when investigating predictors for heart attack.

In this paper, we examine the problem of concept aggregation in the context of a clinical data-mining task. We assess the value of corpus-driven and knowledge-driven methodologies to compute a similarity score for concept pairs. To evaluate similarity within a specific situation we rely heavily on context-specific data. Initial similarity calculations are compiled on a homogenous set of clinical notes, emphasizing the contextually dependent and corpus-driven methodology as a first step. The further refinement of the corpus-based measure is created on two types of ontological knowledge (path length and definitional word overlap), both aiming to differentiate related from semantically similar concept pairs. We evaluate the different methods, including a hybrid score that averages the three measures, on a large dataset of concepts. Our work fits primarily within the field of clinical informatics with the goal of defining a comprehensive way to enrich the analysis of unstructured data located in electronic health records (EHRs).

2. Background

It has been shown that people generally agree upon the notion of similarity or relatedness between ideas [4,5]. As a result, there has been a large effort across various disciplines, including natural language processing [6,7] and biomedical informatics [8–11], to

* Corresponding author. Fax: +1 212 305 3302.
E-mail addresses: rimma@dbmi.columbia.edu (R. Pivovarov), noemie@dbmi.columbia.edu (N. Elhadad).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2012.01.002

create automated methods that can find semantically similar concepts. Much of the research focuses on the identification of both similar and related concepts. Relatedness indicates a semantic association between concepts, such as "ear" and "nose", while similarity specifies that two concepts can be used interchangeably [12]. The focus of this paper is on similarity. Therefore, although many interesting methods have been published on relatedness identification, they are outside the scope of this paper.

2.1. Methods for semantic similarity calculation

Methods developed to identify semantic similarity among concepts fall loosely into two categories – knowledge-based (edge-based and syntactic) and corpora-based (distributional semantics), where information-content-based measures can span both. In this section, we review previous work with specific emphasis on the methods we later use for comparison (and are included in the publicly available UMLS-Similarity package [13]).

2.1.1. Edge-based

Many methods have been developed for a hierarchical interpretation of similarity, based on the location of the concepts in an ontology and the paths among them. Some of the most common methods rely on edge counting, shortest path, and ontological depth [6,14,15], while others add the least common subsumer (LCS) to capture the granularity of a concept in the ontology [16,17]. More recent advances have incorporated into similarity computation the distance to the LCS, assigning weights to the different path types (ontological depth, distance from concepts to LCS) [18], as well as all of the superconcepts between two terms as a way to account for multiple inheritances [19]. We list a few of them below.

Conceptual distance (CDist) [14]

$$\mathrm{sim}_{\mathrm{cdist}}\ (C1, C2) = |\mathrm{shortest\_path}\ (C1, C2)| \tag{1}$$

Leacock and Chodorow (lch) [15]

$$\mathrm{sim}_{\mathrm{lch}}\ (C1, C2) = -\log(|\mathrm{shortest\_path}\ (C1, C2)|/(2 * \mathrm{depth}\ (\mathrm{ontology}))) \tag{2}$$

Wu and Palmer (wup) [16]

$$\mathrm{sim}_{\mathrm{wup}}\ (C1, C2) = 2 * \mathrm{depth}\ (\mathrm{LCS})/(\mathrm{depth}\ (C1) + \mathrm{depth}\ (C2)) \tag{3}$$

Al-Mubaid and Nguyen (nam) [17]

$$\mathrm{sim}_{\mathrm{nam}}\ (C1, C2) = \log((|\mathrm{shortest\_path}\ (C1, C2)| - 1) * (\mathrm{depth}\ (\mathrm{ontology}) - \mathrm{depth}\ (\mathrm{LCS})) + 2) \tag{4}$$

2.1.2. Information-content (IC) based

IC-based methods aim to create measures that incorporate the specificity of a concept within a similarity calculation. The IC calculation is based on the concept and all of its descendants' frequencies within a corpus of texts. The original measure proposed by Resnik evaluated the information shared by two concepts by measuring the IC of their LCS [20]. As the Resnik measure can assign perfect similarity to any two concepts that share the same LCS, two other measures were proposed by Lin [21] and Jiang and Conrath [22]. They also take into account the IC of the concepts themselves, Lin using ratios and Jiang and Conrath using subtraction. More recently, Pirro and Seco devised a similarity measure founded on the idea of "intrinsic IC" which quantifies IC values by relying on the structure of an ontology itself as opposed to a separate corpus [23].

2.1.3. Distributional semantics

Distributional semantics follow the assumption that the meaning of a target word or concept can be acquired from the distribution of words surrounding it, as a whole over its many mentions in a collection of texts. Thus, similarity between two concepts can be quantified according to the amount of overlap between their overall contexts. Here, by context, we are referring to a weighted count of all the words in the sentences surrounding all the instances of a concept. Distributional semantics have been applied to several problems in biomedical informatics [24]. The distributional semantics methodology represents an abstraction of patterns over a larger corpus, where individual mentions of terms are agglomerated to derive an overall pattern of usage. As the abstraction occurs over many mentions and the words in the vocabulary are weighted (typically tf-idf weights), individual negations and other modifiers all contribute to the salient textual patterns present in the corpus. As distributional semantics allow us to compare two concepts in their usage and thus assess their semantic similarity, conversely, such a representation can help perform word sense disambiguation as different senses of a word will appear with different words and phrases surrounding them [24].

The work of Pedersen et al. forms the basis of our context-based similarity measure [11]. Pedersen et al. calculate similarity based on patterns of usage in text with the help of a context vector (which in their case, relied on the Mayo Corpus of Clinical Notes). Each concept of the corpus is represented as a sum of all word vectors that map to the concept, each of dimension the size of the vocabulary. The vector representing word *w* at index *t* is the number of times *w* and *t* co-occur in the same line of a note in the corpus. The similarity between two concepts is then computed as the cosine similarity between their corresponding context vectors. Pedersen found that "the ontology-independent Context Vector measure is at least as effective as other ontology-dependent measures" [11]. Our note-based similarity approach differs mainly in the type of corpus we employ to derive the context vectors. Furthermore, we investigate to which extent this method and ontologically based methods, previously used independently of each other, can be used in complement.

2.1.4. Definitional

The idea of relying on the content of word definitions for assessing appropriate word senses was original proposed by Lesk [25]. The Lesk algorithm selects the sense of a word in a text, which has the highest word overlap between its definition and its context in the text. Banerjee and Pedersen [26] adapted this method further using WordNet and essentially reversed the methodology for the assessment of semantic relatedness (they also added WordNet hyponyms into the computation). Given the Lesk measure, which identifies overlaps in the extended definitions of the two concepts, the relatedness score is defined as the sum of the squares of the consecutive word overlap lengths. A similar methodology was employed by Hamon and Grabar in the biomedical domain [27].

2.1.5. Other methods

Other published measures include similarity calculations between sets of concepts [28], weights of different features in Gene Ontology (GO) [8], and a nonlinear model that is a function of various ontological features such as path length, depth, and local density [29]. Additionally, Rodríguez and Egenhofer [30] focused on hybrid methods that compute both over term definitions and various hierarchical attributes such as features and neighborhoods. Petrakis et al. [31] refined the methodology further to compute neighborhood similarity.