Journal of Biomedical Informatics 45 (2012) 471–481

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.

2.2. Context-aware computing

The notion that the context surrounding information is important is not a novel one and many have thought about applying it in the medical context. Specifically, applications have been developed to facilitate context-aware data mining that would help provide background when evaluating the similarity of the data mining results [32].

Others have devised methods to convert traditional similarity measures into contextually dependent ones. Wu et al. propose a method in which given a similarity measure and a training set they are able to create a new distance function using the ''kernel trick'' [33]. Dong et al. describe a method of ontological conversion designed to take context into account [3]. Other work has looked at the various types of context and how they affect similarity judgments specifically in the case of geospatial IR [34,35]. Most of these context measures are created to enhance personalization across information retrieval systems where the context consists of user, system, and background information. Our method is designed to incorporate an aggregate disease context over many patient records to create disease-specific similarity calculations.

3. Methods

Our composite methodology consists of three complementary similarity measures (Fig. 1). One primary measure is context-based and relies on distributional semantics of patient notes authored by clinicians, while the other two are knowledge-based and rely on concept definitions and their relationships in the Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) ontology. Starting from a homogenous corpus of notes (i.e., notes about patients who share at least one clinical problem), notes are pre-processed to extract concepts mentioned in the corpus. A three-way filter is applied to prune out the extracted concepts and keep a homogeneous set of concepts to be aggregated. The context-based similarity ranks all pairs of concepts. The top-k pairs with the highest context-based similarity are then reordered using the two knowledge-based similarity measures. This section describes the dataset and its pre-processing, the filtering of concepts, the three measures, and the experimental setup for our experiments.

3.1. Data and knowledge sources

Distributional similarity techniques assume that the meaning of a word or concept can be represented by the context in which it appears, across a large number of mentions. As such, the more frequent a concept is in a corpus, the more accurate its context will be at representing the meaning of the concept. Corpus selection is important – a random sample of notes from a random sample of patients might provide a large set of concepts pairs for which to assess similarity, but the concepts might be too sparse and the resulting contexts might be misleading. Our corpus selection process follows.

We chose to collect a homogenous and semantically coherent corpus of clinical notes, in order to ensure that concepts, which are clinically relevant to the patients, are likely to appear frequently enough. For this study, we collected a corpus of notes from patients with chronic kidney disease (CKD). The methods we employ are disease independent, but the fact that we select notes from patients all with at least one condition in common allows us to identify and aggregate concepts frequently mentioned when documenting a particular set of patients. Furthermore, CKD is a prevalent condition in our institution, thus allowing us to collect a large corpus of notes. Patients with CKD have many comorbidities and disorders, providing us with many different concepts to consider in our similarity computation.

Our corpus of clinical notes comes from the NewYork–Presbyterian Hospital (NYPH) Clinical Data Warehouse (CDW). Using ICD-9 codes as evidence of CKD, all notes for CKD patients recorded between 1990 and September 2010 were extracted from the CDW. Each note was processed by our in-house NLP pipeline [36], which identifies document structure (section boundaries and headers, list items, paragraph boundaries and sentence boundaries) [37], performs shallow syntactic analysis (part-of-speech tagging and phrase chunking), and named-entity recognition of UMLS concepts through dictionary matches against the UMLS [1]. The UMLS aggregates terms from different vocabularies and maps them to semantic concepts, each labeled with a Concept Unique Identifier (CUI). The named-entity recognition in our corpus was performed using the 2010AA UMLS version and restricted to the SNOMED-CT terminology. The full pipeline was tested on a manually curated gold standard of 31 notes and yielded an F-measure of 88.55. The pipeline processed a patient note in .26 s on average.

The knowledge-based part of our similarity computation relies on the SNOMED-CT. SNOMED-CT is a terminology of clinical terms and is a primary resource for concept standardization in the clinical domain [38]. SNOMED-CT is particularly useful for our purposes because it provides term definitions and synonyms, as well as semantic relations among concepts. The relationship types have very specifically defined attributes and lend themselves well to our ontological similarity measure. We utilize the concept definitions and synonyms encoded in SNOMED-CT for the definitional similarity. The version of SNOMED-CT we use in this study is from the July 31st 2010 release and consists of over 292,000 active concepts, 760,000 concept descriptions and 824,000 inter-concept relationships.

3.2. Filtration

Given the pre-processed corpus of CKD notes, we can extract a list of all mentioned concepts. In an effort to create a concise and unambiguous list of similar concept pairs, however, we perform a three-tiered filtration step. The filtration relies on the concepts (semantic types), the structure of the notes (section types), and the note category (note types).

The concept filtration follows the hypothesis that two concepts are more likely to be similar if they belong to the same semantic group, whereas two concepts from different semantic groups can

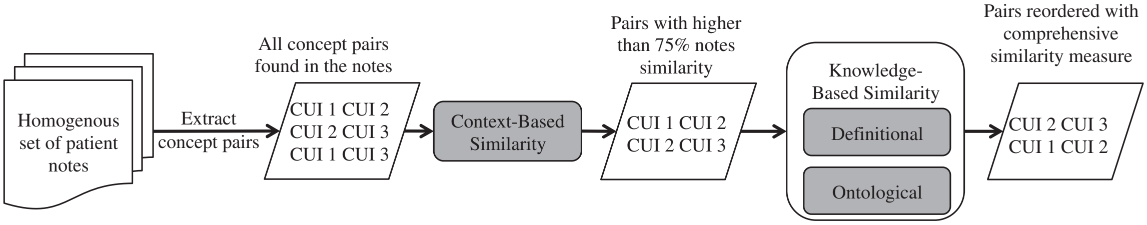
Homogenous set of patient notes
Extract concept pairs
All concept pairs found in the notes
CUI 1 CUI 2
CUI 2 CUI 3
CUI 1 CUI 3
Context-Based Similarity
Pairs with higher than 75% notes similarity
CUI 1 CUI 2
CUI 2 CUI 3
Knowledge-Based Similarity
Definitional
Ontological
Pairs reordered with comprehensive similarity measure
CUI 2 CUI 3
CUI 1 CUI 2

Fig. 1. Our methodology for finding context-dependent similar concepts. An overview of the entire pipeline, beginning with a set of patient notes and ending with ranked list of similar CUI pairs.

indicate semantic relatedness only. For example, it is unlikely that an anatomical concept is similar to a disorder, whereas it is possible that the disorder and the anatomical concept are semantically related (a stroke occurs in the brain, for instance). In practice, limiting the set of semantic types constrains the set of potential relationship types among CUIs, by reducing potential meronyms and focusing instead on hyponyms and metonyms [10]. In this study, we filter in all concept mentions that belong to the Disorder (DISO) semantic group, as defined by McCray et al. [39]. The DISO group contains 12[1] out of the total 133 semantic types. We chose this group as it represents the types of concepts we are interested in examining (e.g., diseases, findings, and signs and symptoms), but the method is agnostic to the chosen semantic group(s).

The section filtration's aim is two fold: to ensure the pool of input concepts is homogenous and to mitigate the potential CUI mapping errors that occurred during pre-processing. To keep the list of input concepts homogenous and specific to the patients under study, we filter out concepts mentioned in the Family History section of the notes. The concepts mentioned in the Medication section are filtered out in an attempt to reduce pre-processing mistakes, arising from the ubiquitous medication abbreviations that are prone to erroneous UMLS mapping. Because our pre-processing pipeline does not perform word-sense disambiguation and keeps all possible CUI mappings instead, the input concept list to our similarity computation can contain incorrect concepts. For instance, "mg", a common abbreviation in the Medication section of notes, which stands for "milligram", is mapped erroneously to "Madagascar", and "Magnesium".

The note filtration operates at a higher level and selects the notes that contain the "richest" content for our purposes. Our initial corpus of CKD notes has more than 1700 different note types (e.g. Primary Provider note, Cardiology Consult note, and Miscellaneous Nursing note). We identified the note types that had more than 60,000 occurrences of SNOMED-CT concepts in the DISO semantic group over the entire set of patients. The filter helps to produce a homogenous, focused list of input concepts, on which to compute pairwise similarities.

3.3. Note-based similarity

The note-based similarity takes two UMLS concepts (CUI) as input and returns a similarity score defined as the cosine similarity between the two concepts' context vectors [40]. The context vectors are derived from the filtered, pre-processed notes for each CUI. They have V elements, where V is the size of the vocabulary. In our experiments, the vocabulary consists of all stemmed words present in the filtered notes.

Given a CUI c and a stemmed word w, the value of the context vector for c at index w is defined as the tf-idf value of w, when c and w occur in the same sentence over the input filtered notes. The tf-idf value is based upon the number of times c and w appear both individually and together. Note that our metric differs from previous work slightly, as we operate over sentences, as opposed to lines in the note [11].

The note-based similarity is computed for all concept pairs. However, because highly infrequent CUIs have very sparse context vectors, which do not represent their context accurately, we only considered CUIs that occur more than five times in the input corpus. Calculation of note-based similarity for all CUI pairs on our corpus took 5.28 h on a linux machine with CentOS 5.4 16-core, 2.93 GHz Xeon X5570 model, 24 GB RAM, with a Hitachi 10,000 RPM drive.

Following our hypothesis that contextual similarity is the basis for semantic similarity, concept pairs with high note-based similarity are kept as candidates for similarity, while all the other pairs are discarded. In our experiments, we set the threshold for note-based similarity at 75%. For instance, in the context of CKD patients, the concepts pairs (Muscle Injury, Traumatic injury of skeletal muscle) and (Acne, Common Acne) both have a cosine similarity above 75%, and are considered for further processing. The concept pairs with lower similarity scores are filtered out, such as (Localized mass, Nodule) and (Chronic Low Back Pain, Pain, NOS) which each have similarity scores in the low 50%.

3.4. Ontological similarity

We describe a novel method for semantic similarity using ontologically defined relationships. We look at the SNOMED-CT ontology as a flat terminology and concentrate on edge types rather than the hierarchy itself. With this view of SNOMED-CT we are able to look at all of the layers that consist of deleted, moved, and retired concepts. This was highlighted as important in the Dong et al. paper that stated: "... in an ontology environment, the types of relations are various, and relations can be defined by multiple restrictions. Obviously, the two factors cannot be ignored when computing similarity for ontology concepts" [3]. The method is based upon the types of relationships between concepts where the different types are broken down into three tiers. The tiers are defined by the characteristic and refinability features of SNOMED-CT relationships (Table 1) and used to group the relationships into ones resembling the most to least closely related. The tiers were used to define weights for various relationships types and the weights were chosen to reflect the strength of each relationship type. There seems to exist a natural hierarchy of relationship strengths that we chose to exploit in the method, such as the observation that non-refinable relationships are of a stronger nature than ones that are optionally or mandatorily refinable. In addition, the weights chosen reflect a system to ensure a score between 0–1 for each relationship type and to delineate between tiers, a twice-larger difference between tiers was introduced (.2) in comparison to the weight difference within each tier (.1). To reduce complications every term listed as both qualifier and defining (Associated Finding, Access, Priority, Clinical Course, Laterality, Associated Procedure, Using Device, Surgical Approach) was grouped in the Defining, Optionally Refinable tier and always given a .5 weight. The weights described here can be tweaked to assess similarity alternatively or even a different semantic relationship; our main contribution lies in proposing a novel way to consider ontological path length calculations.

To find the path between two UMLS CUIs, each CUI was mapped to all of its SNOMED-CT concepts and all possible combinations of pairwise shortest paths were calculated. The average of these paths

Table 1
Relationship Tiers defined by concept Refinability and Characteristic (which are assigned by SNOMED-CT).

Characteristic	Refinability	Example	Tier	Weight
Defining	Not Refinable	Is A	1	1
Historical	Not Refinable	Replaced By	1	.9
Additional	Not Refinable	Part Of	2	.7
Defining	Optionally Refinable	Associated With	3	.5
Qualifier	Optionally Refinable	Measurement Method	3	.4
Qualifier	Mandatory Refinable	Associated Finding	3	.3

[1] Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom.

was taken and assigned as the official path length between the CUIs. The ontological similarity was calculated in 2.18 s for all 794 CUI pairs (15,187 SNOMED concept pairs) on a linux machine with Ubuntu 10.04.03 12-core, 2.4 GHz Opteron 4176 model, 32 GB RAM, with a Dell 7200 RPM drive. The following algorithm (Eq. (5)) was used to assign ontological weights for each individual pairwise path:

$$\mathrm{Sim}_0 = \sum_{e=1}^{E} \mathit{weight}_e/|E| - \alpha(|E| - 1) \tag{5}$$

$E = \{e_1, e_2, \ldots, e_n\}$ where e_i = edge in path, weight_e = assigned weight for edge e, $\alpha = .2$.

For example, one path between C0002622 (Amnesia) and C0751295 (Memory Loss) is between the SNOMED-CT concepts 247606008 and 162199006, with a path similarity of .5333 as illustrated in Fig. 2 and Eq. (6).

$$\mathrm{Sim}_O\ (247606008, 162199006) = (.9 + 1 + .9)/3 - .2(2) = .5333 \tag{6}$$

3.5. Definitional similarity

The third similarity metric we used was definitional similarity. Definitional similarity is a measure of lexical commonality between two concepts – a metric widely used in word sense disambiguation, which can be seen as a reverse goal of our task [26]. We focused on lexical inclusion as the metric and we used the following formula:

$$\mathit{Sim}_D = |(C1 + C2)| - |C1 + C2|/\mathrm{Min}(|C1|, |C2|) \tag{7}$$

where C_i = {words in definition and synonyms of all mappings of CUI "C_i" in SNOMED-CT}.

We chose to use this metric as a way to capture complete subsets as being perfectly similar while adequately capturing the amount of discordance between the two sets of words. While the Lesk methods look at consecutive words, we treat the definitions as a bag of words. For example, the similarity between C0240419: Tender Muscles and C0575064: Muscle Soreness would be between SNOMED-CT concepts 22166009 and 278018006 with a definitional similarity of 1.

22166009 skeletal muscle tender (finding), skeletal muscle tender, muscle tenderness, muscle soreness, tender muscles.

278018006 Tender muscles (finding), tender muscles, tender muscles.

$$C1 = [\text{skeletal}, \text{ muscle}, \text{ tender}, \text{tenderness}, \text{soreness}, \text{muscles}]$$

$$C2 = [\text{tender}, \text{ muscles}]$$

$$\mathrm{Sim}_D\ (22166009, 278018006) = ((6 + 2) - 6)/2 = 1 \tag{8}$$

The definitional similarity calculation for all 794 pairs took .31 s on the same Ubuntu machine used for ontological calculations.

3.6. Experimental setup

Our method finds similar pairs from a very large number of pairs (in our experiments, approximately 14 million); therefore, it would be impossible to create an annotated gold standard list for full evaluation of true negatives and positives. In addition, because the set of input pairs is extracted from a corpus of notes, any gold standard is bound to be corpus-dependent. Therefore, for an evaluation method, we instead assess the accuracy of our methods, its variants and a baseline on a subset of all 14 million pairs, namely the ones with high note-based similarity (i.e., above 75%). In our experiments, there were 794 such pairs.

The evaluation of all three methods was calculated on the 794 pairs (those already filtered by the note-based similarity). The definitional and ontological similarity measures were used and evaluated as secondary metrics. The first evaluation was performed on the note-based method alone to assess its individual contribution. Next, the average of the note-based and ontological methods as well as the average of the note-based and definitional methods were calculated to see the added benefit of each. Finally, the average of all three methods was computed.

To further assess whether a threshold on note-based similarity at the 75% level is appropriate, we calculated similarities and collected gold-standard annotations for 100 random CUI pairs from the 25–50% note-based similarity bracket and 100 random pairs from the 50–75% bracket as well.

3.6.1. Annotations

Two physician annotators evaluated the results of the similarity calculations. The annotators were presented with the 794 CUI pairs in random order along with all of their SNOMED-CT definitions and synonyms. The annotators were specifically asked about the similarity of the concepts within the context of a general population of CKD patients. The annotators were asked to answer the following question with yes, no, or maybe: "Considering a patient with CKD, from a clinical standpoint, would you say that these two concepts could be used interchangeably?" The annotators were not shown any actual medical notes but only a pair of terms. Such an evaluation was chosen as the purpose of our method was to summarize the term usages from the corpus as a whole and use the shared kidney disease framework to find similarity specific to the kidney context overall. The inter-annotator agreement between the physicians was calculated using Cohen's Kappa [41] and after converting all Maybe's to Yes, this resulted in a kappa of .68 without any further adjudication. A kappa of .68 is accepted as representing a substantial amount of agreement between annotators [42]. A conversion of Maybe's to No's resulted in a slightly lower Kappa of .67. The final conversion from Maybe to Yes was appropriate in this instance not only because of the higher Kappa, but also as the "Maybe" was used to annotate terms that are similar in some cases and would require more specific knowledge on the particular patient to determine definitive similarity. A few examples of terms that were marked as "Maybe" are (Swollen Foot,

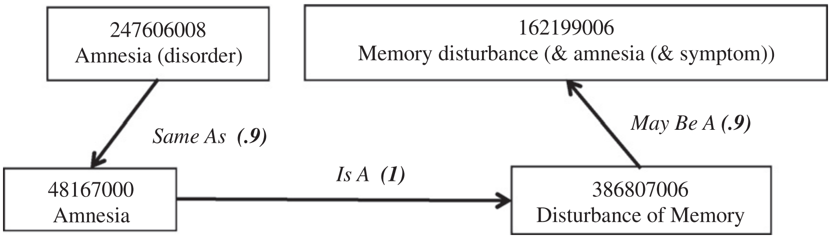

Fig. 2. The shortest SNOMED-CT path between 247606008 and 162199006. An example of a path between two UMLS concepts that were mapped back to SNOMED-CT, the "may be a" edge is not found by hierarchical path methods.

Table 2
All of the UMLS-Similarity measures and their inclusion or exclusion in our baseline.

Category	Measure	Used for Baseline?	Why not?	Parameters
Path	pathcdist	No	Same as the inverse of cdist	–
	lch	Yes	–	SNOMED-CT vocabulary
	nam	Yes	–	Parent/Child and Broader/Narrower
	wup	Yes	–	
		Yes	–	
Information content	jcn	No	IC was calculated on different corpus	–
	lin	No		–
	res	No		–
Definitional	lesk	Yes	–	Definitions from SNOMED-CT vocabulary All relationship types
Note-based	vector	No	Very similar to our vector-based method	–

Swollen Ankle), (Acute and Chronic Inflammation, Focal Chronic Inflammation), (Radiation Burn, Effects of Radiation). The accepting of a "Maybe" annotation as "Yes" gives us more opportunity to create a comprehensive list of similar concepts for large-scale concept aggregation. Given the two annotators' answers for a particular pair, a consensus gold-standard answer was defined as "Yes" if both annotators answered yes, and "No" otherwise. Under this setup, the original list of 794 concept pairs contains 145 pairs annotated as similar.

3.6.2. Baseline

UMLS-Similarity [13] is a Perl package which encodes ten different similarity methods based on ontologies and corpora. UMLS-Similarity encodes path-based measures, information-content measures, definitional measures, and note-based measures. We ran five of the measures on the UMLS-2010AA as baseline, consisting of the path-based and definitional measures, as described in Table 2.

3.6.3. Evaluation metrics

As we are interested in evaluating each similarity measure independently as well as their combined effect, we created receiver operating characteristic (ROC) curves for each combination of similarities at every similarity threshold [43] for both our method and the baseline comparisons. We used the areas under the curves as the single measure to compare how well each similarity measure did.

4. Results

4.1. Dataset

The total dataset collected for this experiment consisted of 2569 patients and their notes, which covered a span of over 20 years (1990–2010). In total, there were 403,819 notes from which we extracted 8869 unique UMLS concepts that are in SNOMED-CT. The minimum unit of computation used in this study was a sentence containing a CUI and the corpus can be described as the set of these sentences. Fig. 3 shows a histogram of the notes with respect to the number of sentences they contain.

4.2. Filtration

The note filtration was used to narrow down the total notes used for the experiment, while keeping rich content. Initially, the corpus consisted of 403,819 total notes with 1739 unique note types. Selecting the note types with more than 60,000 DISO concepts occurrences overall resulted in keeping 17 concept-rich note types (Table 3). This filtration led to a total of 170,775 notes used for the analysis (that is, less than 1% of the note types captures over 40% of the notes). As many institutions have a similar problem of unrestricted note type changes within their medical departments, the rapidly growing and changing EHR note structures make it difficult to keep track of which notes are most important. As the importance of note types is dependent on the question being asked, it is possible to vary the semantic groups included in the analysis, thereby altering the concepts found and the note types considered to be most important. This simple way of ranking note importance by concept density is a dynamic and institution-independent way to identify the most contextually specific salient notes in their EHRs.

4.3. Similarity

To determine the best way to create the context vectors, we performed all experiments with stemming and without stemming the words in the corpus. The stemming approach showed a minor

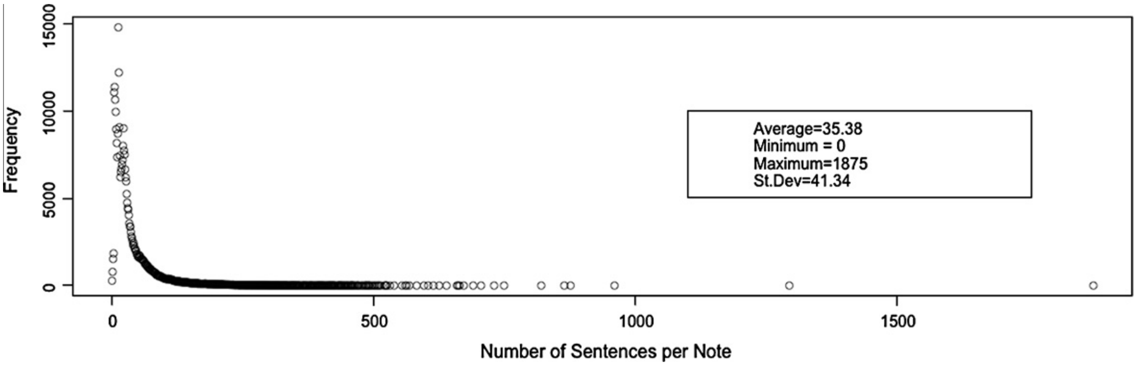

Fig. 3. Descriptive Statistics for sentences in our CKD corpus. This graph illustrates the variety of sentence lengths found and chunked by the ClinNote pipeline.

Table 3
Note types selected through the note filter.

Note type	CUIs that map to SNOMED from the DISO group
Clinical_Note	743268
Discharge_Summary_Note	432138
Admission_Note	341975
Signout	305453
Physical_Therapy	258589
Progress_Note	253971
Nursing_Adult_Admission_History	239946
Surgical_Pathology_Event	223064
Follow-up	204593
Consult_Note	154823
12-Lead_Electrocardiogram	135571
Transthoracic_Echocardiography	112197
Ambulatory_Internal_Medicine_Structured_Note	110324
AMB_Internal_Medicine_Follow-Up_Note	107129
Operative_Note	93521
Primary_Provider_Clinic_Note	91803
X-ray_of_Chest,_Portable	63968

improvement over the unstemmed version and therefore we chose to report the stemmed similarity results.

4.3.1. Experiments for concept pairs with high note-based similarity

We report in this subsection results on a full list of pairs with a note-based similarity above 75% similarity, corresponding to 794 pairs.

The lexicographical comparison of definitions and synonyms of the concepts created a second layer of similarity which we used in addition to the ontological method to move pairs which are simply related to further down on the ranked list than those which are semantically similar. Given the algorithm used for definitional similarity, we often found high similarity between parent–child concepts or concepts with a short definition or list of synonyms.

Fig. 4 shows the ROC curves on the 794 pairs annotated for similarity with different combinations of the similarity measures (combinations represent an average of the individual measures). Although not reported in this paper, we also experimented with the effect of simply applying the ontological and definitional scores to rank and re-order the 794 pairs (without averaging in the note-based score itself). These resulted in slightly smaller AUCs than their note-based averaged counterparts.

Table 4 shows the top-10 concept pairs ranked by the composite method, which averages the three similarity scores for each pair.

Table 4
Top-10 concept pairs found by the composite (average of note-based, ontological-based, and definitional-based) method.

CUI 1	CUI2	Similarity score	Similar?
C1998242 Traumatic injury of skeletal muscle	C0410256 Muscle injury	1	Y
C1691215 Penile hypospadias	C0848558 Hypospadias	0.972	Y
C0240419 Muscle tenderness	C0575064 Skeletal muscle tender	0.966	Y
C2678517 Thrill (finding)	C0232269 Cardiac thrill (finding)	0.958	N
C0677659 Gastro-esophageal reflux disease with esophagitis	C0014869 Peptic esophagitis	0.95	Y
C0149889 Anorectal fistula	C0205929 Anal fistula	0.945	Y
C0158458 Acquired hallux valgus	C0018536 Hallux Valgus	0.937	Y
C0520474 Aseptic Necrosis of Bone	C0029445 Bone necrosis	0.935	Y
C1261287 Stenosis	C0009814 Acquired stenosis	0.935	Y
C0243095 Finding	C0037088 Signs and symptoms	0.933	Y

It is difficult to assess the coverage of our approach to identify similar concept pairs over the original set of 14 million pairs. Instead, we investigate to which extent the automatically discovered candidate pairs are relevant for the input corpus. Since the goal of this method is to aggregate concept pairs that are semantically similar, it is important to know whether the discovered pairs are frequent enough in the input corpus. We assessed the coverage of the concepts that made it into our annotated list of 794 pairs. The concepts made up for 6% of the total number of concepts in the corpus. They are common concepts however, as they cumulatively make up for 30% of the concept frequencies in the corpus. This confirms that our approach is appropriate for discovering pairs of similar concepts, which are frequent in the corpus, and thus important to discover.

We compared our comprehensive method with five methods (lch, wup, cdist, nam, lesk) packaged in the UMLS-Similarity perl

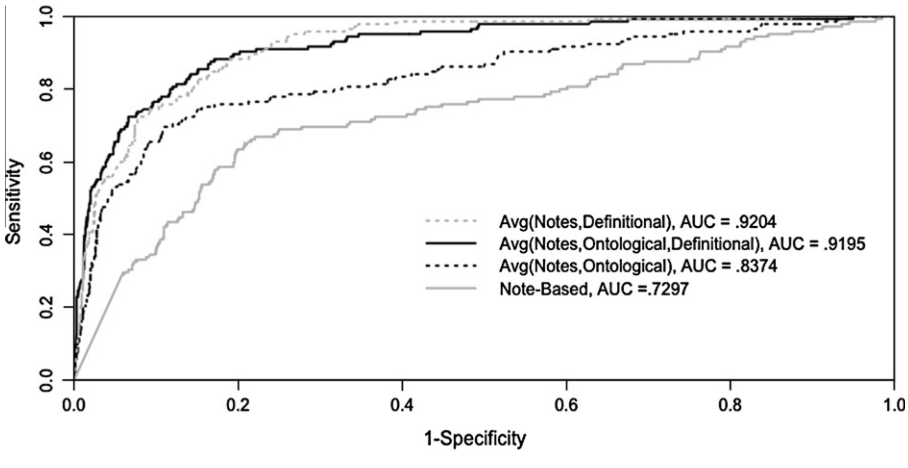

Fig. 4. ROC curves using our methodology. We compared the curves of all similarity combinations between note-based, ontological-based, and definition-based measures.

program. As all of the methods we used except Lesk rely upon the hierarchical relationships present in the UMLS only (Parent/Child or Broader/Narrower), they frequently missed paths between concepts. This happens because many concepts are linked with non-hierarchical relationships such as "moved to" or "deleted from". Table 5 shows the number of missing paths (from the 794 total pairs, 145 of which are similar) when using PAR/CHD, RB/RN, or both. RB/RN misses the largest amount of paths and while PAR/CHD does quite a bit better, even combining the two, leaves over 10% of the total and 20% of the truly similar concept paths as null. We present the ROC curves for the combined PAR/CHD and RB/RN hierarchical methods as well the Lesk method, although the Lesk and CDist scores are coarse whole number measures and lead to fairly few data points for the ROC curves (Fig. 5). The baselines all underperform compared to our three similarity metrics and their combinations.

We compared our final list of similarities with the list of 566 concept pairs collected by Pakhomov et al. [4] in their semantic similarity study. There were only four pairs in common (Rhonchi, Rales), (Polydipsia, Polyuria), (Vertigo, Dizziness), and (Constipation, Diarrhea). Only Rhonchi/Rales were found as similar by the annotators.

4.3.2. Experiments for concept pairs with low note-based similarity

In the experiments described above, we focus on concept pairs with a high note-based similarity (above 75%). This assumes that the context from notes is the most salient cue towards semantic similarity compared to definitional and ontological similarities. As a sanity check that note-based similarity does not miss pairs that are semantically similar, we expanded our gold standard with 200 more concept pairs and collected similarity assessment from our judges, following the same methodology as in the above data set: 100 random pairs from the set of pairs with a note-based similarity between 50% and 75% similarity and 100 random pairs from the set of pairs with a note-based similarity between 25% and 50% similarity. We calculated the path-based similarity and the ontological similarity for the 200 concept pairs. Out of the 200 pairs, only three were scored as interchangeable, and thus similar: (Respiratory alkalosis, Alkalosis), (Disturbance in sleep behavior, Sleep disorders) and (Liver palpable, Liver edge). Furthermore, none of the three were unanimously assigned a "Yes" by our experts. While, this is only for a random sample, it provides face validity to the claim that note-based similarity is the primary factor to assess context-based semantic similarity.

Table 5
Missing paths in hierarchical methods. For the "PAR/CHD or RB/RN" method, we used a PAR/CHD path if it existed and RB/RN path otherwise.

	Broader/ Narrower (RB/RN) relationships (%)	Parent/Child (PAR/ CHD) relationships (%)	PAR/ CHD or RB/RN (%)
Pairs with no path	77.8	15.6	11.3
Pairs with no path that are similar	75.9	24.1	20.0

5. Discussion

5.1. Impact of context

The experiments confirm that context plays a crucial role in assessing similarity of medical concepts: the writing patterns of clinicians provide valuable information to determine which concepts are mentioned in similar contexts and thus are good candidate pairs for aggregation. However, these patterns are all the more visible because the information used for the note-based similarity is derived from a large corpus, with a coherent set of concepts, all related to a particular topic (chronic kidney disease in our experiments). For example, consider the two concepts "Difficulty Hearing" and "Complete Deafness". Generally, the two might be similar enough for aggregation but not given a history of kidney disease. One of our physician annotators pointed out that difficulty hearing might serve as a clue of an adverse drug event caused by an overly high dosage of medication. Complete deafness does not offer the same reaction, as total deafness is rarely an adverse drug event. Such examples illustrate the need for context-dependent similarity measures.

In our experiment, we found the inter-annotator agreement between the physicians was quite high given the subjective nature of the question. The fair amount of agreement indicates that physicians generally concur on medical concept similarity within a particular context. It is a testament to the fact studied generally by Tversky [5] and in the medical arena by Pakhomov et al. [4,44] that there is a universal concept of similarity that most people agree upon.

5.2. Impact of the ontological-based similarity

The relationship-based ontological measure we proposed was able to locate many more paths than other popular methods encoded in the UMLS-Similarity package. Because the baseline methods rely upon hierarchical relationships only, they are often unable

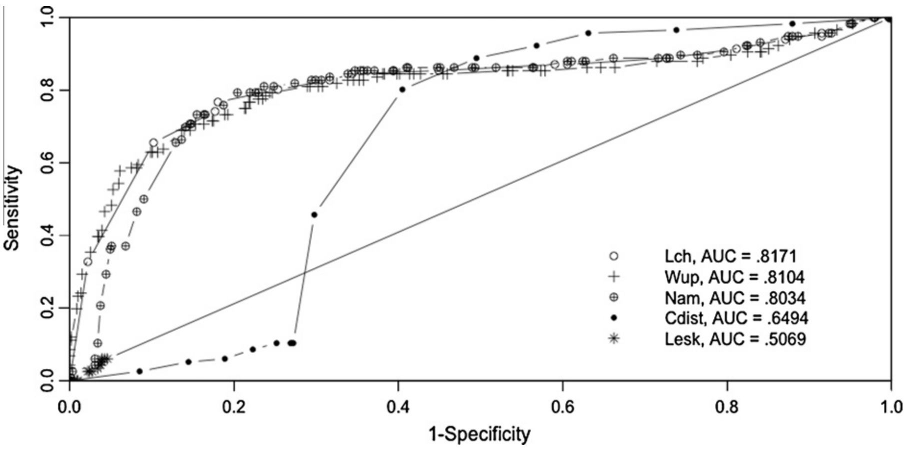

Fig. 5. ROC curves with five methods encoded in UMLS-Similarity. We calculated the ROC curves including only the pairs between which paths were found, leaving only 116 similar pairs and 704 pairs in total.

to find the complicated trajectories among concepts and focus primarily on more straightforward pairings. This limits the types of paths that can be found between two concepts. In contrast, our ontological similarity incorporates all types of ontological relationships into its path computation, and due to the disease/disorder and SNOMED-CT only filtrations, ensures that a path will always be found between two concepts. Our method takes care to assign greater weight to more significant relationships (Is-A), but does incorporate edges between concepts that others do not, such as "may be a" or "moved to".

5.3. Impact of the definitional-based similarity

The definitional similarity measure also provided important information about similarity and did surprisingly well, serving as the best individual similarity measure when averaged with the context-based primary. Nevertheless, it must be emphasized that this measure performs well as a discriminatory measure after the initial contextual similarity thresholding, and would probably not perform well on its own to discover candidate similar pairs. In fact, when applied to the 14 million pairs, 85% of them share no words in their definitions, and thus have a 0.0 definitional-based similarity measure and conversely, 2099 pairs have a perfect 1.0 definitional-based similarity measure. On its own, this measure does not have enough granularity to rank pairs. It performs well combined with more nuanced similarity metrics.

Leveraging the concept definitions for assessing similarity can be viewed as a word sense disambiguation method (the original proposition by Lesk was to use definitions in exactly this way [25]). Given the potential tagging ambiguity that may arise during the entity-recognition phase and would result in a perfect note-based similarity score, the definitional similarity helps to move apart those incorrectly perfectly similar concepts.

5.4. Impact of combining data-driven and knowledge-driven similarity measures

The note-based similarity measure, which relies on patterns of clinicians writing their notes, and the knowledge-based similarity measures, which rely on ontological knowledge, provide complementary cues to the assessment of concept similarity. Concepts that appear far away from each other in the ontology, but are used comparably in the clinical notes are generally irrelevant in the CKD context and can be aggregated for the purpose of data mining. For instance, the concepts C0025874 (Metrorrhagia) and C0312414 (Menstrual spotting) have a low ontological similarity score of .126, but a note-based score of .904, and thus can be correctly identified as similar. Alternatively, concepts may appear very close in a medical ontology but be used vastly differently in context. Such use indicates general similarity but a notable difference in the context of CKD. For example, the concepts C1444079 (Focal chronic inflammation) and C0021376 (Chronic inflammation) had a full ontology-based similarity of 1, but the note-based score was .814. In general, we found that it was essential to incorporate ontological and definitional similarity to separate pairs misguided by abbreviations used in medical text. Although the note-based similarity helps anchor variants of the same concept it has no mechanism for word-sense disambiguation specifically when an abbreviation maps to multiple SNOMED concepts. Out of the 794 pairs that were evaluated, 80 of them had a note-similarity of 1, indicating that they were used in exactly the same way throughout the entire corpus. Most of this is due to a shared "trigger", as our named-entity recognition tool maps each word to all possible UMLS CUI matches. The letters RA could trigger both Rheumatoid Arthritis and Refractory Anemia, giving them a note-based similarity of 1, but clearly they are semantically different. Of the 80 pairs with perfect note similarity, 42 of them were actually similar such as (Incomplete Spontaneous Abortion, Incomplete Abortion Unspecified) both triggered by "Incomplete Abortion" and (Tonsillar Carcinoma, Malignant neoplasm tonsil) both triggered by "Tonsil Cancer"; the rest were disambiguated thanks to the addition of ontological and definitional similarity.

5.5. Limitations

A limitation in our method stems from the heuristic-based weighing of the SNOMED-CT relationships and our incorporation of all relationship types in the method. These ontological artifacts do not fully incorporate the semantic relationship between the concepts, but our method remains robust despite these inherent ontological quirks and our algorithm is able to find the existence of paths where other ontology-based methods are not.

Another limitation of our research is the dependence on the note-based similarity. The ranking and re-ordering performed by definitional and ontological similarity is executed on the pre-selected set of pairs derived from the note-based similarity. We chose to implement a cut-off because although all three similarities can be calculated for all 14 million pairs, there would be no feasible way to collect gold-standard annotations for each pair. As our method is geared towards identifying context-dependent similarity and the similarity measure that incorporates context is note-based, we decided to implement a threshold dependent on this particular similarity. The threshold was set to 75% as a heuristic, by looking at the curve of the similarity values and picking enough to demonstrate the methodology and provide the annotators with a manageable set of pairs to evaluate. Although we demonstrated that this threshold seems to provide a reasonable set of similar terms and prune out non-similar pairs, in the second annotation experiment, it can be adjusted and refined in further studies.

5.6. Future work

When examining the pairs of similar concepts produced by our composite method, we noticed potential for expanding our method to higher dimensions of similarity and clustering concepts. The pairwise similarity often produced triangulated results, which suggest clustering could be carried out as an extension of the pairwise similarity methodology to identify groups of concepts that are semantically similar enough to be aggregated. For instance, we located multiple triplets (three concepts vaguely describing the same concept with each pair achieving high similarities) and one five-pair cluster with the five different concepts describing sputum of different colors (yellow, green, brown, clean, and white). We found each of the 10 combination pairs scoring similar.

We also examined the semantic types of the 707 unique CUIs that make up our final pairings. More than 30% of the CUIs were of the "Disease or Syndrome" semantic type – a larger portion than any other type within the Disorder semantic group. We also compared the semantic types of paired CUIs, almost half of which had identical semantic types. The rest were mostly pairings between the "Finding" and "Disease or Syndrome" types. Similarly, only 2/3 of the 145 pairs annotated as similar were of the same semantic type. The frequent combination of Finding and Disease/Syndrome could help fuel questions of disease etiology and can be explored further with other semantic groupings.

Other methods which leveraged context vectors [11] have used larger corpora consisting of about one million notes and report minimal filtration based on note sections. These approaches are focused on using large corpora to battle the concept sparsity in the data. We found that a methodical filtration to create a cohesive set of notes can help minimize noise created by the inherent nature of medical sublanguage displayed in clinical text. In future work, we may consider defining a direction of similarity as discussed by Kotlerman et al. to incorporate more pairs, pairs with a non-symmetric similarity [45].

We found that CUIs present in the final list of pairs accounted for 30% of the CUI frequencies of the entire corpus. This finding emphasizes how important a homogenous context is to similarity calculation. Additionally, this opens up further research on how CUI occurrences may inherently bias some of the context vector scores even with appropriate tf-idf weighting.

During the initial pre-processing steps, we limit our study to only one semantic group, as we are searching for only similarities but we know there are cases where similarity goes beyond synonymy. A particularly interesting type of relationship is a metonymy, such as the relationship between chemical compounds and lab tests that measure those compounds, for example "Glucose measurement, urine" and "Glycosuria". We plan to investigate further the connection between the semantic type-based filtration and types of relationships discovered.

6. Conclusion

Clinical corpora contain much information waiting to be mined. Mapping clinical term mentions to semantic concepts in an ontology provides valuable abstraction from lexical variants present in text. But some concepts might need to be further aggregated in order to avoid problems of signal dilution. Our approach scores the similarity of two input concepts by combining complementary information derived from usage patterns of clinical documentation, accepted definitions, and position of the concepts in an ontology. Our experiments show that, given a coherent corpus of clinical notes, it is possible to determine automatically which concepts convey similar meaning in the context of the corpus with accuracy above that of previously proposed methods. Finally, this study provides insight into the notions of concept relatedness and similarity, both critical to capturing and representing knowledge in the biomedical field.

Acknowledgments

The authors would like to thank Dr. Herbert Chase, Dr. Hojjat Salmasian, and Dr. Michael Owen for their similarity annotations and numerous discussions. We also thank Mitzi Morris for her advice and help as well as two anonymous reviewers for their valuable comments on earlier versions of this paper. Rimma Pivovarov is supported by the National Library of Medicine predoctoral fellowship 5T15LM007079-19. This study is supported by R01 LM010027 (N.E.) and NLM contract HHSN276201000024C (N.E.).

References

[1] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inform Med 1993;32:281–91.

[2] Popescu M, Xu D. Data mining in biomedicine using ontologies. Massachusetts: Artech House Publishers; 2009.

[3] Dong H, Hussain FK, Chang E. A context-aware semantic similarity model for ontology environments. Concurr Computat: Pract Exp 2010;23:505–24.

[4] Pakhomov SV, McInnes B, Adam B, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. In: Proceedings of American medical informatics association symposium, Washington (DC); 2010. p. 572–6.

[5] Tversky A. Features of similarity. Psycol Rev 1977;84:327–52.

[6] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;9:17–30.

[7] Androutsopoulos I, Malakasiotis P. A survey of paraphrasing and textual entailment methods. J Artif Intell Res 2010;38:135–87.

[8] Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes MD. IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics 2010;11:588–603.

[9] Verspoor K, Dvorkin D, Cohen K. Ontology quality assurance through analysis of term transformations. Bioinformatics 2009;25:77–84.

[10] Elhadad N, Sutaria K. Mining a lexicon of technical terms and lay equivalents. In: Proceedings of the workshop on BioNLP: biological, translational, and clinical language processing. Prague; 2007. p. 49–56.

[11] Pedersen T, Pakhomov SV, Patwardhan S, Chute C. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40:288–99.

[12] Budanitsky A, Hirst G. Evaluating WordNet-based measures of lexical semantic relatedness. Comput Linguist 2005;32:13–47.

[13] McInnes B, Pedersen T, Pakhomov SV. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In: Proceedings of the American medical informatics symposium, San Francisco (CA); November 2009. p. 431–35.

[14] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform 2004;37:77–85.

[15] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C, editor. WordNet: an electronic lexical database. Massachusetts: MIT Press; 1998. p. 265–83.

[16] Wu Z, Palmer M. Verbs semantics and lexical selection, In: 32nd Annual meeting of the association for computational linguistics. Las Cruces (New Mexico): association for, computational linguistics; 1994. p. 133–38.

[17] Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. In: 28th Annual international conference of the IEEE engineering in medicine and biology society. New York (USA): EMBS 2006 IEEE Computer Society; 2006, p. 2713–7.

[18] Matar Y, Egyed-Zsigmond E, Lajmi S. KWSim: concepts similarity measure. In: Proceedings of conférence en recherche d'Information et applications (CORIA08). France; 2008. p. 475–82.

[19] Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 2010;44:118–25.

[20] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Mellish CS, editor. 14th International joint conference on artificial intelligence, IJCAI 1995. Montreal, Quebec, Canada: Morgan Kaufmann Publishes Inc.; 1995. p. 448–53.

[21] Lin D. An information-theoretic definition of similarity. In: Shavlik J, editor. Fifteenth international conference of machine learning, ICML 1998. Madison (Wisconsin, USA): Morgan Kaufmann; 1998. p. 296–304.

[22] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: International conference on research in computational linguistics, ROCLING X, Taipei Taiwan; 1997. p. 19–33.

[23] Pirro G, Seco N. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In: Meersman R, Tari Z, editors. OTM 2008 confederated international conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Berlin/Heidelberg (Monterrey, Mexico): Springer; 2008. p. 1271–88.

[24] Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform 2009;42:390–405.

[25] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on systems documentation. New York, NY, USA: ACM; 1986. p. 24–26.

[26] Banerjee S, Pedersen T. An adapted lesk algorithm for word sense disambiguation using WordNet. In: Proceedings of the 3rd international conference on intelligent text processing and computational linguistics. Mexico City, Mexico; 2002. p. 136–45.

[27] Hamon T, Grabar N. How can the term compositionality be useful for acquiring elementary semantic relations? In: Proceedings of the 6th international conference on advances in natural language processing. Berlin. 2008. p. 181–92.

[28] Cordí V, Lombardi P, Martelli M, Mascardi V. An ontology-based similarity between sets of concepts. In: Proceedings of WOA. Italy; 2005. p. 16–21.

[29] Li Y, Bandar Z, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowledge Data Eng 2003;15:871–82.

[30] Rodríguez M, Egenhofer MJ. Determining semantic similarity among entity classes from different ontologies. IEEE Trans Knowledge Data 2003;15:442–56.

[31] Petrakis EGM, Varelas G, Hiliaoutakis A, Raftopoulou P. Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. In: 4th Workshop on multimedia semantics. Greece; 2006. p. 44–52.

[32] Singh S, Vajirkar P. Context-aware data mining using ontologies. In: Proceedings of the 22nd international conference on conceptual modeling. Japan; 2003. p. 405–18.

[33] Wu G, Chang E, Panda N. Formulating context-dependent similarity functions. In: Proceedings of the 13th annual ACM international conference on multimedia. Chicago; 2005. p. 725–34.

[34] Janowicz K. Kinds of contexts and their impact on semantic similarity measurement. In: 5th IEEE workshop on context modeling and reasoning (CoMoRea 2008) at the 6th IEEE international conference on pervasive computing and communication. Hong Kong; 2008. p. 441–6.

[35] Keßler C, Raubal M, Janowicz K. The effect of context on semantic similarity measurement. In: On the move to meaningful internet systems: OTM workshops. Berlin; 2007. p. 1274–84.

[36] Lipsky-Gorman S, Elhadad N. ClinNote and HealthTermFinder: a pipeline for processing clinical notes. Columbia University Technical Report; 2011.

[37] Li Y, Lipsky-Gorman S, Elhadad N. Section classification in clinical notes using a supervised hidden markov model. In: ACM international health informatics symposium (IHI); 2010. p. 744–50.

[38] SNOMED-CT: Scope Memo. July 2010 international release.

[39] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001;10:216–20.

[40] Manning C, Raghavan P, Schütze H. Introduction to information retrieval. Schütze: Cambridge University Press; 2008 [chapter 6].

[41] Cohen JA. A coefficient of agreement for nominal scales. Educat Psychol Measure; 1960.

[42] Landis RJ, Koch GG. Measurement of observer agreement for categorical data. Biometrics 1997;33:159–74.

[43] Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.

[44] Pakhomov S, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute C. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 2011;44:251–65.

[45] Kotlerman L, Dagan I, Szpektor I, Zhitomirsky-Geffet M. Directional distributional similarity for lexical inference. Nat Language Eng 2010;16:358–89.