Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 288–299

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Measures of semantic similarity and relatedness in the biomedical domain

Ted Pedersen [a,*], Serguei V.S. Pakhomov [b], Siddharth Patwardhan [c], Christopher G. Chute [b]

[a] *Department of Computer Science, 1114 Kirby Drive, University of Minnesota, Duluth, MN 55812, USA*
[b] *Division of Biomedical Informatics, Mayo College of Medicine, Rochester, MN, USA*
[c] *School of Computing, University of Utah, Salt Lake City, UT, USA*

Received 11 April 2006
Available online 10 June 2006

Abstract

Measures of semantic similarity between concepts are widely used in Natural Language Processing. In this article, we show how six existing domain-independent measures can be adapted to the biomedical domain. These measures were originally based on WordNet, an English lexical database of concepts and relations. In this research, we adapt these measures to the SNOMED-CT® ontology of medical concepts. The measures include two path-based measures, and three measures that augment path-based measures with information content statistics from corpora. We also derive a context vector measure based on medical corpora that can be used as a measure of semantic relatedness. These six measures are evaluated against a newly created test bed of 30 medical concept pairs scored by three physicians and nine medical coders. We find that the medical coders and physicians differ in their ratings, and that the context vector measure correlates most closely with the physicians, while the path-based measures and one of the information content measures correlates most closely with the medical coders. We conclude that there is a role both for more flexible measures of relatedness based on information derived from corpora, as well as for measures that rely on existing ontological structures.

Keywords: Semantic similarity; Path based measures; Information content; Context vectors; SNOMED-CT

1. Introduction

Semantic relatedness refers to human judgments of the degree to which a given pair of concepts is related. Studies [1,2] have shown that, surprisingly, most humans agree on the relative semantic relatedness of most pairs of concepts. Measures of relatedness are automatic techniques that attempt to imitate human judgments of relatedness. Many such techniques [3–6] already exist in the realm of domain-independent Natural Language Processing. However, the lack of domain-specific coverage of the resources used by these measures makes them ineffective for use in domain-specific tasks. Because of the availability of numerous ontologies and resources in the biomedical domain, it is possible to adapt these measures and apply them to domain-specific tasks.

The existence of semantic equivalence classes between lexical items in English makes it highly desirable to use thesauri of synonymous or nearly synonymous terms for information retrieval (IR) and document retrieval (DR) applications. The issue is particularly acute in the medical domain due to stringent completeness requirements on such IR tasks as patient cohort identification. We believe that measures of semantic similarity and relatedness can improve the performance of such systems, since they are able to map a user's specific search query to multiple equivalent formulations. For example, a user's query for "congestive heart failure" could be expanded to include the semantically similar terms of *cardiac decompensation*, *pulmonary edema, ischemic cardiomyopathy* and *volume overload*. Clearly, *pulmonary edema* does not denote the same or even a similar disorder as *congestive heart failure* but

[*] Corresponding author. Fax: +1 218 726 8240.
E-mail address: tpederse@d.umn.edu (T. Pedersen).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.06.004

under the patient cohort identification conditions it could be considered as an equivalent search term.

In this research, we adapt a number of measures of similarity and relatedness to the biomedical domain. We should emphasize that semantic *relatedness* and semantic *similarity* are two separate notions. Semantic relatedness is a more general notion of the relatedness of concepts, while similarity is a special case of relatedness that is tied to the likeness (in the shape or form) of the concepts. A measure of semantic similarity takes as input two concepts, and returns a numeric score that quantifies how much they are alike. Such a measure (e.g., [3–5]) is usually based on *is-a* relations found in the underlying taxonomy or ontology in which the concepts reside. For example, *common cold* and *illness* are similar in that a *common cold* is a kind of *illness.* Likewise, *common cold* and *influenza* are similar in that they are both kinds of *illness*. Of course, many ontologies include additional relations between concepts such as *has-part, is-a-way-of-doing, is-a-cause-of, is-a-symptom-of, etc.* that are not directly accounted for in measures of similarity. Thus, we view semantic similarity as a special case of semantic relatedness, and we believe that developing measures that take advantage of increasingly rich ontologies (particularly in the biomedical domain) which have a wealth of relations beyond *is-a* is an important area of future work. This is especially relevant in light of progress in automatically identifying a wide range of semantic relations in medical text (e.g., [7]). We believe that the work in this article provides a necessary stepping stone to tackling the more general problem of identifying semantically related concepts.

Outside the biomedical domain, measures of semantic similarity and relatedness have proven useful in a number of NLP tasks. For example, Budanitsky and Hirst [8] identify malapropisms using various measures of similarity and relatedness. The premise of their approach is that a word that is not semantically similar or related to its neighbors may be a word that is misspelled but is accidentally a valid word, as in “The *nights* of the round table rode again.” Resnik [9], Patwardhan et al. [10], and McCarthy et al. [11] carry out word sense disambiguation based on the idea that a word should be used in the sense that is most similar to or related to the sense of the words that surround it.

However, all of this work has been domain-independent and has been based on WordNet [12], which is a freely available lexical database that represents an ontology of approximately 100,000 general English concepts. In the biomedical domain, there are a growing number of ontologies that organize medical concepts into hierarchies and semantic networks, perhaps best exemplified by the Unified Medical Language System (UMLS®) of the National Library of Medicine (NLM). One of the largest and most extensive sources included in UMLS® is SNOMED-CT®. To date, relatively little work has been done on developing and evaluating measures of conceptual similarity and relatedness for such resources. The premise of this article is that some of the measures that have been found to be effective with WordNet can be adapted and extended to SNOMED-CT®, thereby making it possible to automate certain NLP tasks in the biomedical domain.

This article starts by describing other related work in the biomedical domain as well as domain-independent techniques that implement or use measures of semantic similarity. The article then proceeds by introducing the various resources that we use in measuring similarity and relatedness among concepts. This includes SNOMED-CT®, the Mayo Clinic Corpus of Clinical Notes, and the Mayo Clinic Thesaurus. We then describe the five measures of similarity and one measure of relatedness (all based on WordNet) that we have adapted to the biomedical domain.[1] We also introduce a new test bed for the evaluation of measures of semantic similarity and relatedness in the biomedical domain. Finally, we present our experimental results, and suggestions for future work.

2. Related work

In the biomedical domain, measures of semantic similarity based on ontologies were developed as early as 1989. Rada et al. [13] devised a “semantic distance” measure based on semantic networks. They used MeSH as their semantic network, which consists of biomedical terms organized in a hierarchy. Indeed, one of the measures described in this paper was inspired by this work. Taking a similar approach, Caviedes and Cimino [14] developed the CDist measure for finding path lengths in the UMLS hierarchy. Two of the measures we compare in this paper are path-based measures. Our work with these measures primarily differs in the ontology being used to compute the paths and the path-lengths between concepts.

Recently, Lord et al. [15] adapted WordNet-based measures of relatedness to the Gene Ontology [16], which is a highly specialized ontology of the molecular functions and biological processes of gene products. The hierarchy also describes cellular components associated with gene products. In this work, they find that the semantic similarity of proteins based on the ontology has a high correlation with “sequence similarity,” a separate measure based on the protein sequences. Our work, in comparison, deals with more general biomedical concepts, and provides a more robust evaluation against a manually created data set.

In addition to adapting existing measures to the biomedical domain, there has been some work in creating new techniques for measuring the similarity of terms and concepts. Work by Wilbur and Yang [17] defines a strength metric which is used to retrieve relevant articles using lexical techniques. The metric uses the correlation between the occurrences of a term in documents with the subjects of the documents to define the strength of the term. This metric is

[1] All six of these measures were originally implemented in the WordNet::Similarity package, which can be found at http://search.cpan.org/dist/WordNet-Similarity.

used by the PubMed service to index and retrieve relevant biomedical documents. Research done by Spasic and Ananiadou [18], defines a new similarity metric based on a variation of *edit distance* [19] applied at a word level. In short, the semantic similarity of two terms is the cost associated with converting one term to another, using *insert, delete* and *replace* operations on words (instead of letters). This method additionally uses the UMLS taxonomy to minimize the effect of word variants. It also varies the costs associated with the operations, based on the "semantic load" of the word being edited. For example, deleting a known term present in the UMLS has a higher cost than deleting a conjunction.

Later in this paper, we describe a measure based on context vectors—inspired by Schütze's [20] method for word sense discrimination. This, in turn, is an adaptation of Latent Semantic Indexing [21] commonly used in Information Retrieval. Latent Semantic Indexing (LSI) and Latent Semantic Analysis (LSA) have been shown to be useful in the biomedical domain for indexing and retrieval of clinical records [22,23], for classifying medical events [24,25] and for managing variations in medical terminology [26]. In our research, we use some LSA principles for the vector-based measure.

In the realm of domain-independent Natural Language Processing, semantic similarity has been recently used in numerous tasks such as spelling correction [8], word sense disambiguation [10,27], information extraction [28] and textual inference [29]. All of these applications show that semantic similarity and semantic relatedness have proven useful in a domain-independent setting.

3. Knowledge sources

In this section, we describe the three biomedical information resources used by the measures. All of the measures described later in this article were originally based on WordNet. Since WordNet is a domain-independent lexical resource, it has very little coverage in the biomedical domain (as shown by [30]). To make the measures more effective in a domain-specific setting, we substituted the underlying domain-independent knowledge sources with resources from the biomedical domain. A description of these resources follows.

3.1. SNOMED-CT®

SNOMED-CT® (**S**ystematized **No**menclature of **Med**icine, **C**linical **T**erms) is an ontological/terminological resource that has a wide coverage of the clinical domain. It is produced by the College of American Pathologists and is now distributed as part of the UMLS® through the National Library of Medicine. SNOMED-CT® is used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerized physician order entry, disease surveillance, image indexing and consumer health information services. The version of SNOMED-CT® we use in this study is from 2004 and consists of more than 361,800 unique concepts with over 975,000 descriptions (entry terms) [31].

The terminology is organized into 13 hierarchies at the top level: clinical findings, procedures, observable entities, body structures, organisms, substances, physical objects, physical forces, events, geographical environments, social contexts, context-dependent categories, and staging and scales. There is one overarching root node that joins all 13 hierarchies together. The concepts and their descriptions are linked with approximately 1.47 million semantic relationships including *is-a, assists, treats, prevents, associated etiology, associated morphology, has property, has specimen, associated topography, has object, has manifestation, associated with, classifies, has ingredient, mapped to, mapped from, measures, clinically associated with, used by, anatomic structure is physical part of.*

3.2. The Mayo Clinic Corpus of Clinical Notes

This resource consists of ~1,000,000 clinical notes collected over the year 2003 which cover a variety of major medical specialties at the Mayo Clinic. Clinical notes have a number of specific characteristics that are not found in other types of discourse, such as news articles or even scientific medical articles found in MEDLINE. Clinical notes are generated in the process of treating a patient at a clinic and contain the record of the patient-physician encounter. These notes are typically dictated and represent a kind of quasi-spontaneous discourse [32] where the dictations are made partly from notes and partly from memory. More often than not, the speech tends to be telegraphic which presents certain challenges for Natural Language Processing.

At the Mayo Clinic, the dictations are transcribed by trained personnel and are stored in the patient's Electronic Medical Record. These transcriptions are then made available for health science research. The notes are semi-structured where each note consists of a number of subsections such as Chief Complaint (CC), History of Present Illness (HPI), Impression/Report/Plan (IP), Final Diagnoses (DX), among others.

We are particularly interested in the CC, HPI, IP and DX section of the clinical notes. The CC section records the reason for visit; HPI section has information of what other treatments/problems the patient has had in the past; IP section contains the diagnostic and current treatment information, while the DX section is a summary of the IP section – it contains only a list of diagnoses. Other sections such as SI (Special Instructions) and CM (Current Medications) are less interesting from the standpoint of semantic relatedness measures, although if we were to focus on computing semantic relatedness between medications, then we may want to consider the CM section as well. The SI section contains administrative information that is not relevant to the patient's condition. The CM section contains a list of medications that were prescribed to the patient. The medications on this list may or may not be related to the condition described in the note; however, the relevant medications tend to be repeated in the IP and HPI sections. We have eliminated the CM section from consideration for now because it may introduce spurious associations on the one hand and may be redundant with the IP and HPI sections on the other.

3.3. The Mayo Clinic Thesaurus

The Mayo Clinic Thesaurus is a rich source of clinical problem descriptions that have been systematically collected at the Mayo Clinic since 1909. The Mayo Clinic Thesaurus has its roots in the Plummer indexing system, introduced at Mayo to index clinical problem descriptions in 1909; it was implemented using 5×8 inch index cards. The index was substantially modified to a bi-axial nomenclature in the course of migrating the Mayo index to IBM Hollerith cards around 1935, and expanded again during our migration to electronic computing environments from 1960. Since 1996, short summaries of patient diagnoses have been created, manually coded and stored in a database. At the time of this study, this resource contained over 16 million unique diagnostic phrases expressed through natural language that are classified into over 21,000 diagnostic categories and represents an utterance level thesaurus. The 16 million phrase-category pairs contain 5,167,428 unique phrases that represent diagnostic statements. Each diagnostic statement has been recorded by a practicing physician at the Mayo Clinic as part of the patient's medical record, manually coded and cataloged for subsequent retrieval using a Mayo Clinic modified Hospital International Classification of Diseases Adaptation (HICDA). The HICDA classification is a hierarchy consisting of four levels. The top level is the most general and has 19 categories such as *Neoplasms, Diseases of the Circulatory System,* etc. The next three levels group diagnoses into more specific categories.

The Mayo Clinic Thesaurus is constructed on the assumption that if several diagnostic phrases have been classified to the same category in the HICDA hierarchy, then these phrases can be considered as synonymous at the level of granularity afforded by HICDA. For example, diagnostic phrases such as "primary localized hilar cholangiocarcinoma" and "cholangiocarcinoma of the Klatskin variety" are linked in a thesaurus-like fashion because these two statements have been manually classified the same way. We consider these two phrases nearly synonymous and use them to generate quasi-definitions for terms found in both SNOMED-CT® and this utterance level thesaurus of diagnostic phrases.

We attempt to reduce the inevitable noise and redundancy in this collection by excluding those phrases that occur 5 times or less and those phrases that are classified as "*Admission, diagnosis not given.*" After these restrictions, the original 5,167,428 diagnostic statements are reduced to a vocabulary of 381,673 terms. Of these, 9951 (2.6%) are also found on the list of SNOMED-CT® descriptions via simple string matching. The terms were placed in lowercase prior to matching.[2] After lowercasing, the list of descriptions from SNOMED-CT® contained 798,168 unique terms. The overlap of 9951 terms with the Mayo Clinic Thesaurus constitutes 1.3% of the total number of unique lowercased SNOMED-CT® terms. Due to the simplicity of the matching method, these data provide only a very rough approximation to the actual intersection between the three vocabularies. These data do suggest however that incorporating SNOMED-CT® into the Mayo Clinic Thesaurus would increase the coverage of either terminology taken separately.

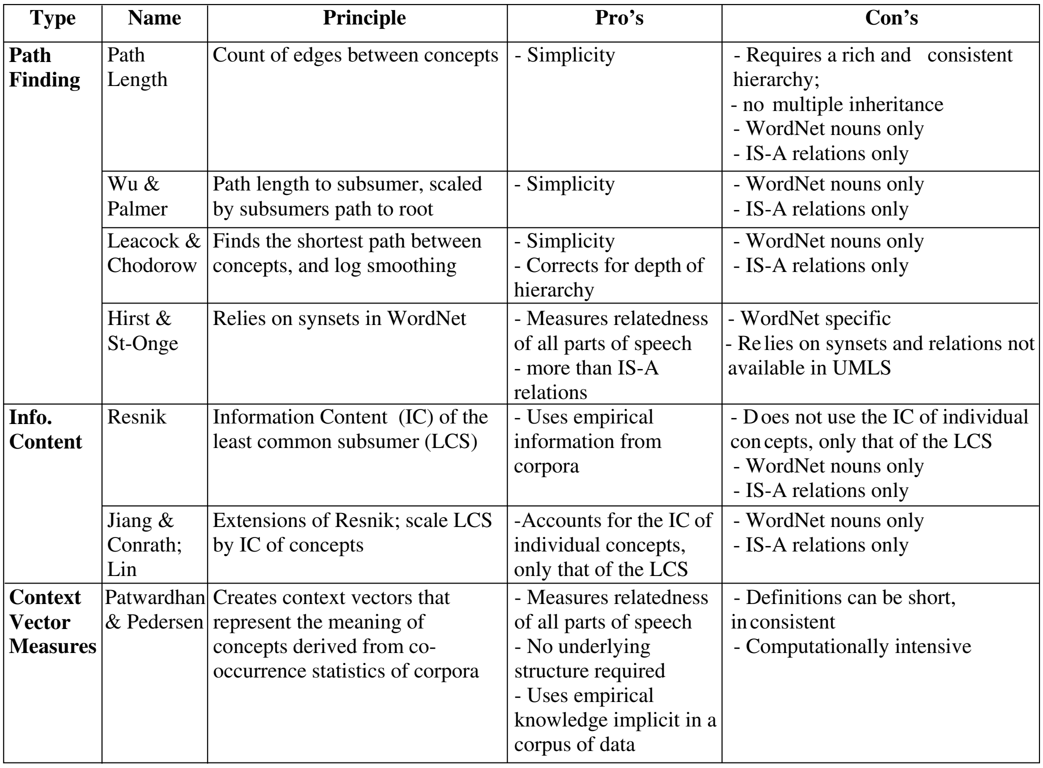
The Mayo Clinic Thesaurus is augmented by merging it with the Medical Subject Headings (MeSH) sub-hierarchy of the UMLS® (version 2003AB) and the 2003 version of SNOMED-CT® (prior to its incorporation into the UMLS®). This augmentation allows mapping from SNOMED-CT® concepts to clusters of terms in the Mayo Clinic Thesaurus. If a term appeared in more than one source, the duplicates were eliminated after the terms were linked to a Mayo Clinic Thesaurus ID.

4. Measures of semantic similarity and relatedness

Semantic relatedness refers to the judgments by humans regarding the relatedness of pairs of concepts. Humans usually agree of the relative relatedness of concepts [1,2]. For example, most humans would agree that *bird* is more related to *feather* than it is to *fork* or to *car*. Research [2,33,34] has shown that humans use the context of words and concepts to build a mental semantic representation of concepts. Over time humans encounter similar contexts for different concepts. Consequently, humans tend to agree on the semantic relatedness of concepts.

Many ideas have been proposed to automatically calculate the semantic relatedness of words to correspond closely to those by human subjects. Some of the commonly used methods derive statistical information from text corpora and combine that information with a lexical resource such as WordNet to make semantic relatedness judgments that have been shown to have a high correlation with those of human subjects [8,10]. Additionally, these techniques have been shown to be useful for many Natural Language Processing tasks such as word sense disambiguation [10,27], spelling correction [8] and information extraction [28]. In this section, we describe several measures based on WordNet that attempt to quantify the semantic relatedness of concepts. Additionally, we

[2] Simple lowercasing was used to maximize the processing speed. A more sophisticated approach to lexical normalization such as the Lexical Variant Generator (LVG) developed at the National Library of Medicine would result in improved matching.



Type	Name	Principle	Pro's	Con's
Path Finding	Path Length	Count of edges between concepts	- Simplicity	- Requires a rich and consistent hierarchy; - no multiple inheritance - WordNet nouns only - IS-A relations only
	Wu & Palmer	Path length to subsumer, scaled by subsumers path to root	- Simplicity	- WordNet nouns only - IS-A relations only
	Leacock & Chodorow	Finds the shortest path between concepts, and log smoothing	- Simplicity - Corrects for depth of hierarchy	- WordNet nouns only - IS-A relations only
	Hirst & St-Onge	Relies on synsets in WordNet	- Measures relatedness of all parts of speech - more than IS-A relations	- WordNet specific - Relies on synsets and relations not available in UMLS
Info. Content	Resnik	Information Content (IC) of the least common subsumer (LCS)	- Uses empirical information from corpora	- Does not use the IC of individual concepts, only that of the LCS - WordNet nouns only - IS-A relations only
	Jiang & Conrath; Lin	Extensions of Resnik; scale LCS by IC of concepts	-Accounts for the IC of individual concepts, only that of the LCS	- WordNet nouns only - IS-A relations only
Context Vector Measures	Patwardhan & Pedersen	Creates context vectors that represent the meaning of concepts derived from co-occurrence statistics of corpora	- Measures relatedness of all parts of speech - No underlying structure required - Uses empirical knowledge implicit in a corpus of data	- Definitions can be short, inconsistent - Computationally intensive

Fig. 1. Classification of measures of semantic similarity and relatedness and their relative advantages and disadvantages.

describe how these measures were adapted to make more accurate judgments in the biomedical domain. The five measures of semantic similarity all use SNOMED-CT®, while the three similarity measures based on information content also use the Mayo Clinic Corpus of Clinical Notes. The Context Vector measure of relatedness only uses the Mayo Clinic Corpus and the Mayo Clinic Thesaurus.

Before describing the measures, we would like to emphasize the difference between *semantic similarity* and *semantic relatedness*. Semantically similar concepts are deemed to be related on the basis of their likeness. Semantic relatedness, on the other hand, is a more general notion of relatedness, not specifically tied to the shape or form of the concept. In other words, semantic similarity can be considered a special case of semantic relatedness. The measures of semantic similarity described here are based on *is-a* relations that link concepts (directly or indirectly) found in a hierarchy. These measures can simply be based on the path lengths between concepts, or they may augment such structural information with corpus based statistics. Measures of semantic relatedness are more general, and can include information about other relations, or may be based on co-occurrence statistics from corpora. We describe several existing measures of similarity and relatedness in this section, focusing particularly on those that we adapted for use with SNOMED-CT®. A general classification of the measures and their relative advantages and disadvantages is provided in Fig. 1.

4.1. Path finding measures

When concepts are organized in a hierarchy, where more general concepts are near the root of the hierarchy, and more specific ones near at the leaves, it is convenient to measure similarity according to the path lengths between concepts. In fact, there have been a variety of such approaches proposed in both the biomedical domain and in domain-independent NLP techniques.

Rada et al. [13] developed a measure based on path lengths between concepts in the Medical Subject Headings (MeSH) ontology, which is distributed by the National Library of Medicine. They relied on *broader than* relations, which link to successively more or less specific concepts as you travel from concept to concept. They used this measure to improve information retrieval by ranking documents retrieved from MEDLINE, a corpus made up of abstracts of biomedical journal articles. More recently, Caviedes and Cimino [14] developed a measure called CDist which finds the shortest path between two concepts in the UMLS®. Their evaluation relative to a small set of concepts and concept clusters drawn from a subset of the UMLS® (consisting of MeSH, ICD-9-CM[3] and SNOMED-CT®) shows that even such relatively simple approaches tend to yield reliable results.

Wu and Palmer [35] present a measure of similarity for general English that relies on finding the most specific con-

[3] International Classification of Diseases, 9th revision, Clinical Modification.

cept that subsumes both of the concepts being measured. The path length from this shared concept to the root of the ontology is scaled by the sum of the distances of the concepts to the subsuming concept. Leacock and Chodorow [36] define a similarity measure that is based on finding the shortest path between two concepts and scaling that value by twice the maximum depth of the hierarchy, and then taking the logarithm of the resulting score. In both of these measures, the path length between concepts is scaled in some way by the overall depth or size of the hierarchy, to avoid reliance strictly on path lengths, which can be misleading due to the fact that the semantic similarity between concepts that is represented by a single link will vary depending on where that link is found in the hierarchy. A link between two very general concepts may imply a reasonably large difference between the concepts, while one between two very specific concepts might represent a small difference.

There have been relatively few attempts to develop path based measures that rely on relations beyond *is-a.* Given the richness of relations found in resources such as SNOMED-CT®, we believe that this is a promising area of future work, however, we have not included such measures in this study. One example of a possible candidate for adaptation is the relatedness measure of Hirst and St-Onge [37]. Their measure, which is based on WordNet determines the relatedness between two concepts by finding the nature of the path that joins them. A path that is not too long and has relatively few changes in direction represents a relatively higher degree of relatedness as compared to a path which is long with many changes in direction.

For the experiments in this article, we developed two path-based measures: a path-length measure for SNOMED-CT®, and an adaptation to SNOMED-CT® of a measure proposed by Leacock and Chodorow. The path-length measure essentially computes the similarity between two concepts by counting the numbers of nodes on the shortest path between them in SNOMED-CT®'s *is-a* hierarchy. The shortest path includes both the concept nodes. The inverse of the path length is defined as the similarity of the two concepts. The adaptation of the Leacock and Chodorow measure is very similar to the path-length measure, except that the Leacock and Chodorow measure scales this shortest path length by the depth of the taxonomy. Mathematically, the similarity of two concepts c_1 and c_2 using the path-length measure (*path*) is defined as:

$$sim_{path}(c_1, c_2) = 1/p, \tag{1}$$

where p is number of nodes on the shortest path between the two concepts in SNOMED-CT®. Similarly, the similarity of two concepts c_1 and c_2 using the Leacock and Chodorow measure (*lch*) is computed as

$$sim_{lch}(c_1, c_2) = -\log\left(\frac{p}{2 \cdot depth}\right), \tag{2}$$

where p is number of nodes on the shortest path between the two concepts in SNOMED-CT® and *depth* is the maximum depth of the hierarchy.

Note that SNOMED-CT® allows multiple inheritance, i.e. a node in the hierarchy can have multiple parents (possibly in different parts of the taxonomy). Thus multiple possible paths can exist between any two concepts. However, we select only the shortest path among those, for both the measures.

4.2. Information content measures

The limitation of purely path based measures is that the degree of semantic similarity implied by a single link is not consistent. Links found between very general concepts convey somewhat smaller amounts of similarity than do links between very specific concepts.

Resnik [3] attempts to address this problem by augmenting concepts with a corpus-based statistic known as *information content*, which is essentially a measure of the specificity of a concept. The information content of each concept in a hierarchy is calculated based on the frequency of occurrence of that concept in a large corpus of text. A concept with high information content is very specific, while lower information content values are associated with more general concepts.

The information content of a concept is estimated by counting the frequency of that concept in a large corpus of text. Note, however, that a single concept can be mapped to multiple lexical terms in text, and conversely a single lexical term can be mapped to multiple concepts. Thus, to get frequency estimates for concepts, Resnik suggests distributing the frequency count of a term equally over the concepts it maps to. But it has also been shown [10] that assigning all of the concepts mapped to a single term, the same frequency count as the term also works well. In this research, the frequency count assigned to a concept is the sum of the frequency counts of all the terms that map to the concept.

Additionally, the definition of information content requires that the frequency count of every concept include the frequency counts of all subsumed concepts in an *is-a* hierarchy. For example, the frequency count for the concept of *disease* would include frequency counts of *tuberculosis* and *influenza* (among others). Similarly, the concept corresponding to the root node of the *is-a* hierarchy has the maximum frequency count, since it includes the frequency counts of every other concept in the hierarchy. Thus, the frequency counts associated with concepts higher up in the *is-a* hierarchy is always greater than or equal to those lower down in the hierarchy.

After obtaining the frequency counts of all concepts, the information content of each concept c is computed as:

$$IC(c) = -\log\left(\frac{freq(c)}{freq(root)}\right), \tag{3}$$

where $freq\,(c)$ is the frequency of concept c and $freq\,(root)$ is the frequency of the root of the hierarchy.

Using this notion of information content, Resnik [3] defines a measure of similarity that holds that two concepts are semantically similar proportional to the amount of information they share. The quantity of shared information is determined by the information content of the most specific concept in the hierarchy that subsumes both the given concepts, which is referred to as the Lowest Common Subsumer. Mathematically, the Resnik measure (*res*) computes the similarity of concepts c_1 and c_2 as:

$$sim_{res}(c_1, c_2) = IC(lcs(c_1, c_2)), \tag{4}$$

where $lcs(c_1,c_2)$ is the lowest common subsumer of concepts c_1 and c_2, and *IC* returns the information content of the concept.

However, the Resnik measure may not be able to make fine grained distinctions since many concepts may share the same Lowest Common Subsumer, and would therefore have identical values of similarity. Jiang and Conrath [5] and Lin [4] developed measures that scale the information content of the subsuming concept by the information content of the individual concepts. Lin does this via a ratio, and Jiang and Conrath with a difference. The Jiang and Conrath (*jcn*) measure computes the semantic distance (inverse of similarity) of concepts c_1 and c_2 as:

$$dist_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \cdot IC(lcs(c_1, c_2)) \tag{5}$$

and the Lin measure (*lin*) computes semantic similarity of concepts c_1 and c_2 as:

$$sim_{lin}(c_1, c_2) = \frac{2 \cdot IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}, \tag{6}$$

where $lcs(c_1, c_2)$ is the lowest common subsumer of concepts c_1 and c_2, and *IC* returns the information content of the concept.

In our study, we have adapted all three of these measures (*res*, *lin* and *jcn*) to the biomedical domain by using the the *is-a* hierarchy of SNOMED-CT®. We used the Mayo Clinic Corpus of Clinical Notes as the source of the frequency counts of the SNOMED-CT® concepts, required to derive their information content values.

4.3. Context Vector measure

Patwardhan [6,38] developed a measure of semantic relatedness that represents a concept as a Context Vector. This is intended to be a more general representation than similarity measurements, since the source of the information for the context vectors is a raw corpus of text, not the paths found between concepts in an ontology. This technique is an adaptation of Schütze's [20] method of word sense discrimination, which is in turn an adaptation of Latent Semantic Indexing [21] as practiced in Information Retrieval. In this technique, we build co-occurrence vectors that represent the contextual profile of concepts. The cosine of the angle between vectors corresponding to two given concepts then determines the relatedness of those concepts.

We start by creating *Word Vectors,* which are first order context vectors, for every content word *w* in our corpus of text. The dimensions of these vectors are content words from the same corpus of text (each dimension corresponding to one content word). The vector for a word *w* is created as follows:

1. Initialize the first order context vector to a zero vector $\vec{w}$.
2. Find every occurrence of word *w* in the given corpus.
3. For each occurrence of *w*, increment by 1 those dimensions of $\vec{w}$ which correspond to the words present in a specified window of context around *w*.

The first order context vector $\vec{w}$, therefore, encodes the co-occurrence information of word *w* and is called its *word vector*. In this research, we use the Mayo Clinic Corpus of Clinical Notes to create word vectors for all content words occurring in the clinical notes. We used one line of text as the window of context.

Having created a set of word vectors, we then use these to create *context vectors* corresponding to every SNOMED-CT® concept whose frequency in the corpus of clinical notes exceeds a predefined threshold. We use the Mayo Clinic Thesaurus to get a list of *descriptor terms* for each concept. The word vectors corresponding to the descriptor terms of a concept are then aggregated to get the context vector for that concept. Thus, a SNOMED-CT® concept is represented as the resultant of descriptor term word vectors, where word vector represents the "contextual profile" of the term, as computed from the Mayo Corpus of Clinical Notes. For example, the SNOMED-CT® concept *"angina pectoris"* (SNOMED-CT ID: 367416001) maps to a cluster of terms in the Mayo Clinic Thesaurus (Cluster ID: M00587016). This cluster also contains terms such as *"vasospastic angina," "CAD with exertional angina," "angina functional class 2," "ischemic heart disease with angina pectoris"* that are not originally part of SNOMED-CT® but are associated with a SNOMED-CT® concept via the Mayo Clinic Thesaurus. Thus, the context vector for *"angina pectoris"* is computed as the resultant of the word vectors for *"vasospastic angina," "CAD with exertional angina," "angina functional class 2"* and *"ischemic heart disease with angina pectoris."*

The semantic relatedness of two concepts c_1 and c_2 is then computed as the cosine of the angle between their context vectors:

$$rel_{vector}(c_1, c_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|v_1| \cdot |v_2|}, \tag{7}$$

where $\vec{v}_1$ and $\vec{v}_2$ are the context vectors corresponding to c_1 and c_2, respectively.

5. Experimental data

Measures of semantic similarity and relatedness can be evaluated both directly and indirectly. The direct method compares the results of the measures relative to human judgments; common standards for domain-independent

English are provided by pairs of manually rated concepts as created by Rubenstein and Goodenough [1] and Miller and Charles [2]. The indirect methods evaluate measures based upon the performance of an application that relies on the measures. Spelling correction [8] and word sense disambiguation [10] have both been used as applications to evaluate measures. Both of these studies found the similarity measure of Jiang and Conrath to be most effective at improving the results of their application, although the word sense disambiguation evaluation also reported that a measure based on finding overlaps in the definitions of words was equally successful.

There are no existing sets of words for the biomedical domain that have been manually scored for similarity by human experts that could be used as a direct means of evaluation. In this research, we created a test bed of pairs of medical terms that were scored by human experts according to their relatedness. A Mayo Clinic physician (Alexander Ruggieri, MD) trained in Medical Informatics, followed the methodology of Rubenstein and Goodenough and generated a set of 120 term pairs that consists of 30 pairs in each of four broad categories of relatedness values from not related at all (1) to very closely related (4). Subsequently, we had 13 medical coders annotate each pair with a relatedness value on a scale of 1–10. A wider scale was chosen for experimental purposes—it is easier to collapse a wider scale than to expand a narrow one. We later collapsed this scale to match Rubenstein and Goodenough's.

A group of medical coders specially trained to classify clinical diagnoses using the same HICDA classification system as the one used to construct the Mayo Clinic Thesaurus annotated the test set for this study. These medical coders had between five and 14 years of coding experience at the time of the study. Although they do not have formal training in medicine, by virtue of working with clinical records and terminologies have had significant exposure to medical language and we considered them as good candidates for this annotation task.

We implemented two measures of semantic similarity based strictly on the *is-a* relations as found in SNOMED-CT®: the path-length measure and the Leacock and Chodorow measure. We also implemented three measures that are based on a combination of information content statistics derived from the entire Mayo Clinic Corpus of Clinical Notes, and the *is-a* relations provided by SNOMED-CT®. Finally, we implemented a Context Vector measure by finding co-occurrence vectors from the Mayo Clinic Corpus of Clinical Notes based on the descriptor terms associated with concepts in the Mayo Clinic thesaurus. As such the Context Vector measure was the only measure that was not dependent on a hierarchical terminology or ontology in some way. We computed the vectors in two different ways, first using the entire Mayo Corpus of Clinical Notes, and then using just the IP sections.

6. Experimental results

6.1. Inter-annotator agreement

As a control, we had 10 of the 13 medical coders[4] annotate the 30 domain-independent English term pairs in the tests sets of Rubenstein and Goodenough and of Miller and Charles using a 10 point scale. This was done to make sure the medical coders understood the instructions and the notion of relatedness. The correlation of the medical coders' judgments with those of the annotators used by Rubenstein and Goodenough was a relatively high value of 0.84. Similarly, the correlation with the Miller and Charles's test set was 0.88. The correlation on the medical test set of 120 concept pairs was 0.51. To derive a more reliable test set we extracted only those pairs whose agreement was high. This resulted in a set of 30 concept pairs (displayed in Table 1) that were then annotated by three physicians and a subset of 9 medical coders from the 13 who annotated the original 120 pairs. All three physicians are specialists in the area of rheumatology. The fact that all of them specialize in the same sub-field of medicine can be helpful in getting good inter-rater agreement. Each pair was annotated on a 4 point scale: practically synonymous (4.0), related (3.0), marginally related (2.0) and unrelated (1.0). We have listed the term pairs and the averaged scores assigned by the physicians and the experts in Table 1. Term pair 20 (shown in boldface) has been excluded from the test bed because the term "lung infiltrates" was not found in the SNOMED-CT® terminology. Thus, the resulting test set consists of 29 pairs; however, we were able to calculate the inter-rater agreement using all 30 pairs. The average correlation between physicians is 0.68. The average correlation between medical coders is 0.78. We also computed the correlation across the two groups after we averaged the scores each member of the respective groups had assigned to each pair in the test set. The correlation across groups is 0.85.

6.2. Comparison among measures

We scored each of the 29 test bed pairs using each of the measures, and then computed the correlation between the measures' output and the human expert judgment scores shown in Table 1. These correlations are shown in Table 2. The highest correlation is achieved by the Context Vector measure when it is derived only from the IP section of the clinical notes. This is particularly true in the case of correlation with the physician judgments. The choice of corpora to use with Context Vector is clearly critical, since the correlation attained by this measure when using the entire Mayo Clinic Corpus of Clinical Notes drops considerably.

[4] Not all of the experts were always available to us at all times, so the number of annotators changed from one set of annotations to the next. No new experts were added, only subtracted based on their availability and work load.

Table 1
Test bed of 30 medical term pairs; sorted in order of the averaged physicians' scores

Term 1	Term 2	Physician	Coder
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3
Chronic obstructive pulmonary disease	Lung infiltrates	2.3	1.9
Rheumatoid arthritis	Lupus	2.0	1.1
Brain tumor	Intracranial hemorrhage	2.0	1.3
Carpel tunnel syndrome	Osteoarthritis	2.0	1.1
Diabetes mellitus	Hypertension	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Cortisone	Total knee replacement	1.7	1.0
Pulmonary embolus	Myocardial infarction	1.7	1.2
Pulmonary fibrosis	Lung cancer	1.7	1.4
Cholangiocarcinoma	Colonoscopy	1.3	1.0
Lymphoid hyperplasia	Laryngeal cancer	1.3	1.0
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Rectal polyp	Aorta	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Depression	Cellulites	1.0	1.0
Varicose vein	Entire knee meniscus	1.0	1.0
Hyperlidpidemia	Metastasis	1.0	1.0

The scores represent an averaged relatedness value (scale is 1–4) over all participating physician and coder annotators.

Table 2
Comparison of correlations across measures for physicians and coders separately and combined

Measure	Phys.	Coder	Both
Vector (IP only, 1M notes)	0.84	0.75	0.76
Vector (All sect, 1M notes)	0.62	0.68	0.69
Lin	0.60	0.75	0.69
Jiang and Conrath	0.45	0.62	0.55
Resnik	0.45	0.62	0.55
Path	0.36	0.51	0.48
Leacock and Chodorow	0.35	0.50	0.47

Two types of the vector measure are presented—one trained on only the Impression/Report/Plan section of clinical note and the other trained on all sections.

In fact, the Context Vector measure based on all the clinical notes performs much like the Lin measure.

We also note that the Context Vector measure produces a much closer correlation with physicians than with the medical coders. For all other measures, this is reversed. We hypothesize that this is due to the nature of the professional training and activities of the two groups—medical coders are trained in the use of hierarchical classifications, while physicians are trained to diagnose and treat patients. One possible indication from this observation is that the data contained in the clinical notes may reflect certain kinds of semantic relations between medical concepts in the mind of a physician better than a hand-crafted medical ontology such as SNOMED-CT®. By all means, more experimentation is necessary to test this hypothesis.

The three information content measures occupied the middle range of performance, with Lin showing a considerably higher level of correlation to be physicians and medical coders. Both Jiang and Conrath, and Resnik performed at somewhat lower levels than Lin, and at identical levels to each other. This is in contrast to direct evaluations made to the Miller and Charles test set in a domain-independent setting, where both Budanitsky and Hirst [8] and Patwardhan et al. [10] report that Jiang and Conrath achieves much higher levels of correlation than Lin or Resnik.

The overall success of the vector measure suggests that an ontology-independent measure can perform at least as well or better than ontology-based measures. However, in the same way that the Context Vector measure is strongly affected by the corpora it is derived from, the same may be true of the information content measures. An important avenue for future work is to experiment with using different portions of the clinical notes and different types of corpora in arriving at information content estimates.

6.3. Impact of size and type of corpora on Context Vector measure

The Context Vector measure is the most flexible of the measures presented here, and as such requires that a number of informed choices be made in order for it to function effectively. Among the most critical is the amount and type of corpora to derive the vectors. To determine if the section types from the clinical notes had an impact on the performance of the Context Vector measure, we experimented with the four section types when using a 100K word portion of the clinical notes. Table 3 displays the correlation of the Context Vectors derived from these different sections, sorted in order of correlation with the physicians.

The best correlation is achieved on the corpus compiled from the IP sections, closely followed by DX. This is not surprising as the IP section contains the diagnostic information pertinent to the patient's condition and intuitively should contain more closely related terms than other sections. The DX section is a summary of the IP section in that it only contains the diagnoses without additional descriptions. It is interesting to note that each of the subsections result in better performance than context vectors derived from the entire corpus Table 4.

To evaluate the impact of the size of the corpora on the Context Vector measure, we ran experiments with the Context Vector measure on varying amounts of the Mayo Clinic Corpus of Clinical Notes ranging from 100,000 to 1 million words. For these experiments, we used data from four sections of the clinical notes—Chief Complaint

Table 3
Correlation of the Context Vector measure derived from different sections of a 100K portion of the Mayo Clinic Corpus of Clinical Notes

Section	Physicians	Coders	Both	# Tokens
IP	0.56	0.59	0.60	10,883,117
DX	0.53	0.55	0.56	490,417
CC	0.47	0.53	0.53	956,438
HPI	0.46	0.54	0.56	7,487,209
ALL	0.41	0.53	0.51	21,593,156

Table 4
Descriptive statistics (size and overall number of tokens) for varying sized portions of Mayo Clinic Corpus of Clinical Notes

# of Notes	Matrix size	# of Tokens
100K	32594 × 32594	21,593,156
200K	43179 × 43179	43,459,602
300K	50928 × 50928	66,176,995
400K	57328 × 57328	88,885,380
500K	62733 × 62733	111,019,453
600K	64910 × 64910	133,719,224
700K	67382 × 67382	156,467,734
800K	69883 × 69883	179,114,059
900K	72911 × 72911	206,489,197
1000K	75195 × 75195	232,080,038

(CC), History of Present Illness (HPI), Impression/Plan (IP) and Final Diagnosis (DX).

The number of words in each sized portion of the clinical notes is shown in Table 3. The number of words (tokens) is found by excluding all words that occur less than five times and more than 1000 times. The matrix size indicates the number of unique words that are found given those cutoffs. For example, there are 32,594 distinct word types found in the 100,000 word portion of the clinical notes. The co-occurrence matrix used in the Context Vector measure is symmetric, meaning that a word vector is created for every word in the corpus that occurs in the designated frequency range, and two words are said to co-occur with each other when they are on the same line of text.

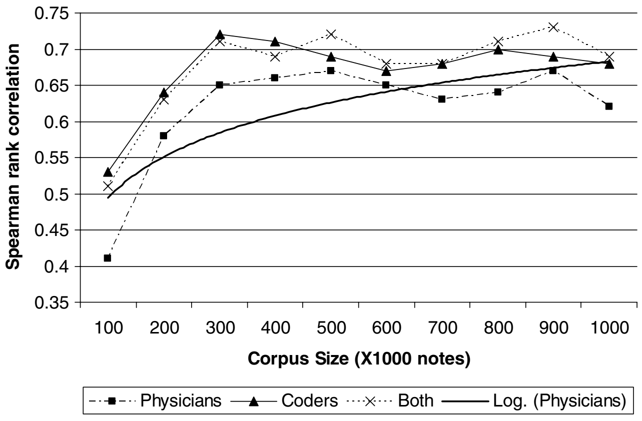


Fig. 2. Correlation of Context Vector measure with human experts across different training corpus sizes. The trendline is fitted to the results obtained by physicians on a logarithmic scale.

Test results relative to the test set of 29 term pairs are shown in Fig. 2. The overall trend suggests that the correlation between relatedness judgments of the Context Vector measure and those of human experts improves with larger amounts of data, where 300K size appears to be the point where gains level off. Fig. 2 shows the correlation of the Context Vector measure with the scores of the physicians and medical coders separately, as well as combined scores averaged across both groups. The log line shows the overall improvement with corpus size.

7. Limitations

Certain limitations of this work must be mentioned to facilitate the interpretation of the results. The main limitation of this research is the relatively low inter-annotator agreement on the initial set of 120 term pairs we compiled to create the test set. To address this limitation, we focused on the main goal of our project which was to compare several established measures of similarity and relatedness that are based on manually compiled ontological knowledge sources to an ontology-independent method. In the context of these comparisons, we felt it was justified to take a subset of 30 of the 120 pairs that were agreed upon by the majority of the annotators. Thus, this smaller set is more reliable but is clearly biased towards "easy" term pairs. The correlation values reported in this paper cannot be interpreted in absolute terms; however, the reduced test set can be used to establish relative performance of different measures.

8. Conclusions and future work

In this article, we have shown the efficacy of adapting measures of semantic similarity and semantic relatedness measures developed for domain-independent English to a specialized subdomain of biomedicine represented by SNOMED-CT®. We have also shown that the ontology-independent Context Vector measure is at least as effective other ontology-dependent measures, provided that there is a large enough corpus of unlabeled training data available. This finding is important because developing specialized ontologies such as WordNet, SNOMED-CT® or UMLS® is a very labor intensive process. Also, there are some indications that manually constructed ontology may not fully reflect the reality of semantic relationships in the mind of a practicing physician. The vector based measure can help alleviate these problems in addition to the benefit of rapid adaptation to a new domain.

In the near future, we plan to extend the measures of relatedness to use the UMLS® as a source of the ontological relations for path-based measures, and to use the UMLS definitions for the context vector measure. We also would like to experiment with applications of semantic relatedness measures to NLP tasks such as word-sense discrimination, information retrieval and spelling correction, in the biomedical domain.

Acknowledgments

We thank the Mayo Clinic Medical Index staff as well as Dr. Alexander Ruggieri, Dr. Peter Kent and Dr. Auethavekiat Paranee for their contribution to the annotation of the test pairs. Dr. Ted Pedersen's role in this work has been partially supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784). This work was also supported by the NLM Training Grant in Medical Informatics (T15 LM07041-19) and the NIH Roadmap Multidisciplinary Clinical Research Career Development Award Grant (K12/NICHD)-HD49078.

References

[1] Rubenstein H, Goodenough J. Contextual correlates of synonymy. Communications of the ACM 1965;8:627–33.

[2] Miller G, Charles W. Contextual correlates of semantic similarity. Language and Cognitive Processes 1991;6(1):1–28.

[3] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence. Montreal, Canada; 1995. p. 448–53.

[4] Lin D. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Madison, WI; 1998. p. 296–304.

[5] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th international conference on research in computational linguistics, Taipei, Taiwan; 1997. p. 19–33.

[6] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 workshop, making sense of sense: Bringing computational linguistics and psycholinguistics together. Trento, Italy; 2006. p. 1–8.

[7] Rosario B, Hearst M. Classifying semantic relations in bioscience texts. In: Proceedings of the 42nd annual meeting of the association for computational linguistics. Barcelona, Spain; 2004. p. 430–7.

[8] Budanitsky A, Hirst G. Semantic distance in WordNet: an experimental application oriented evaluation of five measures. In: Proceedings of the NACCL 2001 Workshop: on WordNet and other lexical resources: Applications, extensions, and customizations. Pittsburgh, PA; 2001. p. 29–34.

[9] Resnik P. WordNet and class-based probabilities. In: Fellbaum C, editor. WordNet: An electronic lexical database. Cambridge, MA: MIT Press; 1998. p. 239–63.

[10] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational linguistics. Mexico City, Mexico; 2003. p. 241–57.

[11] McCarthy D, Keoling R, Weeds J, Carroll J. Finding predominant word senses in untagged text. In: Proceedings of the 42nd meeting of the association for computational linguistics. Barcelona, Spain; 2004. p. 276–86.

[12] Fellbaum C, editor. WordNet: An electronic lexical database. Cambridge, MA: MIT Press; 1998.

[13] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. In: IEEE transactions on systems, man and cybernetics, 1989;19(1): p. 17–30.

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

[15] Lord P, Stevens R, Brass A, Goble C. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics 2003;19(10): 1275–83.

[16] The Gene Ontology Consortium, Gene Ontology: Tool for the Unification of Biology. Nat Genet 2000;25:25–9.

[17] Wilbur W, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Biol Med 1996;26:209–22.

[18] Spasic I, Ananiadou S. A flexible measure of contextual similarity for biomedical terms, Pacific Biocomputing Symposium 2005;10:197–208.

[19] Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 1966;10:707–10.

[20] Schütze H. Automatic word sense discrimination. Comput Linguist 1998;24(1):97–123.

[21] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41:391–407.

[22] Chute C. Classification and retrieval of patient records using natural language: an experimental application of latent semantic analysis. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society. Orlando, FL; 1991. p. 1162–3.

[23] Chute C, Yang Y. An evaluation of concept-based latent semantic indexing for clinical information retrieval. In: Proceedings of the 16th annual symposium on computer applications in medical care. Baltimore, MD; 1992. p. 639–43.

[24] Chute C. The classification of medical events using latent semantic analysis. In: Advances in classification research vol. 2: Proceedings of the second ASIS SIG/CR workshop on classification research. Medford, NJ; 1992. p. 45–51.

[25] Chute C, Yang Y, Evans D. Latent semantic indexing of medical diagnoses using UMLS semantic structures. In: Proceedings of the 15th annual symposium on computer applications in medical care. New York City, NY; 1991. p. 185–9.

[26] Evans D, Chute C, Handerson S, Yang Y, Monardch I, Hersh W. Latent semantics as a basis for managing variation in medical terminologies. In: Proceedings of the seventh world congress on medical informatics (MEDINFO '92). Geneva, Switzerland; 1992. p. 1462–8.

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

[28] Stevenson M, Greenwood M. A semantic approach to IE pattern induction. In: Proceedings of the 43rd annual meeting of the association for computational linguistics. Ann Arbor, MI; 2005. p. 379–86.

[29] Raina R, Ng A, Manning C. Robust textual inference via learning and abductive reasoning. In: Proceedings of the Twentieth national conference on artificial intelligence. Pittsburgh, PA; 2005. p. 1099–105.

[30] Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In: Proceedings of the NAACL 2001 Workshop: WordNet and other lexical resources: Applications, extensions and customizations. Pittsburgh, PA; 2001. p. 77–82.

[31] SNOMED-CT: Fact sheet. 2004. College of American Pathologists.

[32] Pakhomov S. Modeling filled pauses in medical dictations. In: Proceedings of the 37th annual meeting of the association for computational linguistics. College Park, MD; 1999. p. 619–24.

[33] Carnine D, Kameenui E, Coyle G. Utilization of contextual information in determining the meaning of unfamiliar words. Read Res Quart 1984;19:188–204.

[34] McDonald S, Ramscar M. Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. In: Proceedings of the 23rd annual conference of the cognitive science society. Edinburgh, Scotland; 2001. p. 611–6.

[35] Wu Z, Palmer M. Verb semantics and lexical selection. In: Proceedings of the 32nd annual meeting of the association for computational linguistics. Las Cruces, NM; 1994. p. 133–8.

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

[37] Hirst G, St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum C, editor. WordNet: An electronic lexical database. Cambridge, MA: MIT Press; 1998. p. 305–21.

[38] Patwardhan, S. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master of Science Thesis, Duluth, MN: Department of Computer Science. Duluth: University of Minnesota; 2003.