Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet

Ying Liu
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
liux0395@umn.edu

Bridget T. McInnes
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
bthomson@umn.edu

Ted Pedersen
Dept. of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

Genevieve Melton-Meaux
Institute for Health Informatics
University of Minnesota
Minneapolis, MN 55455 USA
gmelton@umn.edu

Serguei Pakhomov
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
pakh0002@umn.edu

ABSTRACT
Automated measures of semantic relatedness are important for effectively processing medical data for a variety of tasks such as information retrieval and natural language processing. In this paper, we present a context vector approach that can compute the semantic relatedness between any pair of concepts in the Unified Medical Language System (UMLS). Our approach has been developed on a corpus of inpatient clinical reports. We use 430 pairs of clinical concepts manually rated for semantic relatedness as the reference standard. The experiments demonstrate that incorporating a combination of the UMLS and WordNet definitions can improve the semantic relatedness. The paper also shows that second order co-occurrence vector measure is a more effective approach than path-based methods for semantic relatedness.

Categories and Subject Descriptors
I.2 [**Artificial Intelligence**]: Natural Language Processing; I.2.7 [**Natural Language Processing**]: Text Analysis; H.3.3 [Information Search and Retrieval]: Clustering

General Terms
Experimentation

Keywords
UMLS, WordNet, Computational Linguistics, Semantic Relatedness.

1. INTRODUCTION
Humans can judge if one pair of concepts is more related than another. Creating computer programs that can do this automatically is our goal. In natural language processing, semantic relatedness measures have broad applications. Ponzetto and Strube [26] used the Wikipedia categorization system as a semantic network to compute the semantic relatedness of words. Chen et al. [7] applied the similarity in machine translation. Other applications include detection of noun phase conjuncts [11], selectional preferences [8] and automatic speech recognition [27]. Semantic relatedness measures can also be successfully used for semantic searching of textual resources available for bioinformatics research. Guo et al. [9] used similarity measures derived from the Gene Ontology for identifying direct and indirect protein interactions within human regulatory pathways. Both Resnik [29] and Patwardhan et al. [22, 23] showed that these various measures can be used to perform word sense disambiguation. Other examples include Bousquet et al [4], who explored the use of semantic distance (the inverse of similarity) for coding of medical diagnoses and adverse drug reactions.

Most existing relatedness judgments are based on knowledge sources such as concept hierarchies or ontologies. For general English text, research on measuring relatedness has relied on WordNet, a freely available dictionary that can also be viewed as a semantic network. For clinical and biomedical vocabularies, they are compiled into the Unified Medical Language System (UMLS) [3], a large lexical and semantic ontology of medical terms maintained by the National Library of Medicine. As the amount of text increases in biomedicine, similarity measures based on hierarchical ontologies are at a disadvantage because they rely heavily upon the structure of the ontology and they are not adaptable to the changes in medical knowledge.

In this paper, we present an ontology-independent semantic relatedness measure that uses second order co-occurrence vectors. The main idea of the measure is that related concepts have a similar context. For example, *doctor* is more related to *stethoscope* than to *telescope* because *doctor* and *stethoscope* share the same medical context. This method takes advantage of large corpora of medical texts, the UMLS Metathesaurus and Semantic Network, and concept definitions from WordNet. It can compute the strength of semantic relatedness between any pair of concepts in the UMLS. In this paper, we use the terms "context

IHI'12, January 28–30, 2012, Miami, Florida, USA.

vector" or "vector" to refer specifically to the second order co-occurrence vector.

The aim of the study is also to determine if the UMLS and other non-medical resources, such as WordNet, contain complementary information. While the UMLS is a very rich source of information on medical and biological terms and concepts, it does not provide full coverage of non-medical concepts, terms and relations [6, 11, 12, and 20]. In this paper, we show that combining the UMLS and WordNet for the purpose of computing semantic relatedness is beneficial.

Our method is an extension of a context vector approach described by Patwardhan and Pedersen [22]. In previous work, a gloss vector was constructed from the WordNet dictionary. We extend the construction of concept definitions to the biomedical domain by using different relations in the UMLS. Our results show that the ontology-independent vector method performs better than the ontology-dependent methods.

2. SIMILARITY AND RELATEDNESS MEASURES

Methods for computing semantic similarity and relatedness are a class of computational techniques. These techniques can be used to create groups of related terms automatically by using information from large corpora and existing ontologies. We treat semantic relatedness as a distinct and more general notion than semantic similarity [10, 21].

2.1 Ontology-Dependent Measures

Ontology-dependent measures of semantic relatedness are based on the ontological relations including is-a, has-part, and is-a-part-of. This dependency on ontological relations can be a disadvantage because ontologies tend to be static and cannot keep up with the rapidly changing structure of knowledge in a given discipline such as biomedicine. One of the simplest similarity measures (path) is a technique based on the calculation of the reciprocal of the shortest path between a pair of concepts in the ontology [28]. This approach inverts the edge counts between two concepts resulting in a similarity score. For instance, in Figure 1, the similarity of *cardiologist* and *doctor* is the same as *pulmonologist* and *doctor* because the distance between them is the same.

Resnik [29] introduced the Information Content (IC) method which associates probabilities to each concept from statistics in a large corpora of text ($IC(c) = -\log p(c)$). The IC value of a concept is estimated by counting the frequency of that concept in a large corpus, along with the frequency of all the concepts that are subordinate to it in the hierarchy. The semantic similarity between two concepts is proportional to the amount of information they share [13, 16, 29]. According to Resnik, the similarity of *cardiologist* and *orthodontist* is the IC value of their least common subsumer (LCS) *doctor*. If two concepts do not share a LCS, the similarity is zero. These methods heavily rely on the ontology structure and were designed primarily for the is-a relation.

More discriminating path-based methods were developed by Wu and Palmer [32] and Leacock and Chodorow [14], which incorporate the depth of the LCS in the ontology in addition to path length. The improvements of Resnik are Lin [16] and Jiang & Conrath [13]. They incorporated the IC measure of every concept. An alternative semantic distance method integrates distributional information and ontological knowledge within a network flow formalism [31]. In our example, *cardiologist-doctor* and *orthodontist-doctor* would have the same similarity score. The IC method has the limitation that concepts at the same level that share the same LCS have the same similarity score.

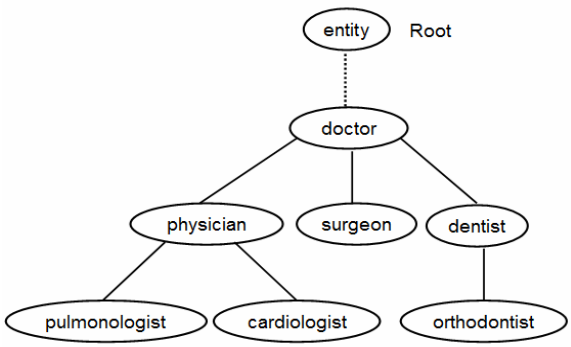


Figure 1. An example of a small portion of a general English hierarchy of concepts.

The UMLS, however, has 12 different types of hierarchical and non-hierarchical relations. The current state-of-the-art in measures of relatedness for medical text are adaptations of existing WordNet-based measures or some variations optimized for specific tasks such as MEDLINE document retrieval or Gene Ontology (GO) searching [17]. Therefore, standard ontology-dependent methods require extensive adaptation to work on the UMLS.

2.2 Ontology-Independent Measures

Ontology-independent methods rely on distributional properties of concepts in large text corpora and may be easier to keep current with changes in a given knowledge domain. These approaches include the semantic relatedness measures proposed by Lesk [15] and Pedersen et al. [24]. Lesk-type approaches calculate the strength of association between a pair of concepts as a function of the overlap between their definitions (Formula (1)). If the definition for *cardiologist* is "a physician who studies the heart" and *orthodontist* is "a dentist who makes teeth straight", there are no overlaps between the two definitions after removing stop words.

$$rel_{lesk} = \sum_{i=1}^{n} freq_{overlap_i} * length^2_{overlap_i} \qquad (1)$$

Banerjee and Pedersen [1] introduced the Extended Gloss Overlap measure which expands the definition by augmenting it with the definitions of senses that are directly connected to it in WordNet. This is often referred to as Adapted Lesk. *Cardiologist* and *orthodontist* are somewhat related in that *cardiologist is* a type of *physicians* and *orthodontist* is a type of *dentists*. This is not apparent in the definition overlaps of *cardiologist* and *orthodontist* but when considering that both *physician* and *dentist* are *doctors*, they are related. This allows for concepts that are indirectly related to be identified and scored by selecting appropriate sources of definitions. The main limitation of a Lesk-type method is that it is based strictly on definitions and does not use any other knowledge source. The definition matching is done by exact word matching. So, two words, such as "doctor" and "physician", occurring in two definitions would not match at all under Lesk. But if they occur in similar contexts they might have some similarity score as assigned by vector. In our experiments,

we use the Lesk method as the baseline to compare with our vector-based method.

To address the limitation in Lesk, Patwardhan and Pedersen [22] introduced the gloss vector measure which combines the definitions of concepts with co-occurrence data. Every word in the definition is replaced by its context vector from the co-occurrence data and relatedness is calculated as the cosine of the angle between the two vectors. For the above example, doctor and physician are replaced by context vectors of the co-occurrence matrix. This avoids the direct matching problem. The advantages of the gloss vector measure are that empirical knowledge implicit in a corpus of data is used, and there is no underlying structure required. The limitation is that the definitions can be short and inconsistent. In this paper, we extend this previous work and apply the second order context vector method to the biomedical domain.

3. EXPERIMENTAL DATA

The described algorithm requires two types of data to calculate semantic relatedness: the text corpus and the concept definitions. Clinical reports, MEDLINE abstracts and general English texts were used as a corpus. The definitions come from the UMLS and WordNet[1].

3.1 UMLS

The Unified Medical Language System (UMLS) is a knowledge representation framework designed to support biomedical and clinical research. It is a widely used database of biomedical terminologies for encoding information contained in electronic medical records and medical decision support. It includes over 100 terminologies and classification systems. The UMLS contains more than 1.7 million active concepts with unique meanings. The three major components of the UMLS are the Metathesaurus, Semantic Network and SPECIALIST Lexicon.

This work focuses on the Metathesaurus which semi-automatically integrates information about biomedical and health-related concepts from various biomedical and clinical sources. Some example terminologies contained in the UMLS include National Cancer Institute Thesaurus (NCIT), SNOMED Clinical Terms (SNOMED CT), and Medical Subject Headings (MSH). The UMLS uses 12 different types of hierarchical and non-hierarchical relations between concepts. The hierarchical relations consist of the *parent*/*child* and *broader*/*narrower* relations. This paper uses only the hierarchical relations to construct the concept definitions. In this study, we limit the scope to UMLS2010AB.

3.2 WordNet

WordNet is a large lexical database of English. Although it includes a certain number of medical terms, a study done by Bodenreider and Burgun [2] showed that the concept overlap between WordNet and the UMLS varies from 48% to 97%. This is because the UMLS records the variability of the lexical forms encountered in the source vocabularies, while WordNet only records the canonical forms. We use a WordNet 3.0 for the current study.

3.3 Text Corpus – Clinical Reports

The inpatient clinical reports were collected from 2003 to 2008 at Fairview Health Services. Located in Minneapolis MN, Fairview Health Services is a non-profit, academic health system that partners with the University of Minnesota. These semi-structured reports consist of admission history, physical operation, discharge summaries, and consultation notes. In raw form, these reports contain on average 500 words. After pre-processing of the text including removal of stop words, numerals and punctuation, each note contained approximately 300 words. Thus the total size of the clinical reports corpus used in this study was ~208.7 million words.

Table 1. The composition the clinical reports

	number of records	number of words (million)	size (MB)
admission	196,778	57.5	496
discharge	305,249	67	591
operation	362,358	76.6	667
consultation	11,504	7.6	66

3.4 Text Corpus – MEDLINE Abstract

MEDLINE contains over 20 million biomedical articles from 1966 to the present. The database includes journal articles from almost every field of biomedicine. For the current study we used article abstracts as the corpus to build a term-term co-occurrences matrix for subsequent computation of semantic relatedness. We used the 2010 MEDLINE abstract[2].

3.5 Text Corpus – English Gigaword

The English Gigaword[3] corpus is a newswire archive that contains international news articles and is maintained and distributed by the Linguistic Data Consortium. The co-occurrence matrix uses approximately 1 million articles (1.3G) from the Xinhua News Agency.

3.6 Reference Standard

The reference standard used in our experiments was based upon a set of medical pairs of terms created specifically for testing automated measures of semantic relatedness as part of a different study [21]. The pairs of terms were compiled by first selecting all concepts from the UMLS with one of three semantic types: disorders, symptoms and drugs. Subsequently, only concepts with entry terms containing at least one single-word term were further selected for potential differences in similarity and relatedness responses. Five medical residents (1 woman and 4 men; mean age 30) at the University of Minnesota Medical School were invited to participate in this study for a modest monetary compensation. They were presented with 724 medical pairs of terms on a touch sensitive computer screen and were asked to indicate the degree of relatedness between terms on a continuous scale by touching a touch sensitive bar at the bottom of the screen. The overall inter-rater agreement on this dataset was moderate (Intraclass Correlation Coefficient - 0.50); however, we were able to select a subset of the ratings consisting of 430 pairs with good agreement (Intraclass Correlation Coefficient - 0.73) and the distribution of ratings and semantic types similar to the original set.

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

[2] MEDLINE abstract are from National Library of Medicine: http://mbr.nlm.nih.gov/Download/index.shtml

[3] English Gigaword are from Linguistic Data Consortium: http://www.ldc.upenn.edu/DataSheets/

4. METHODS

For a pair of concepts with definitions, the basic idea of the second order co-occurrence vector method is to compare the two definitions. Instead of direct matching, such as the Lesk [15] method, our approach finds the context distribution of every word in the definition based on a co-occurrence matrix constructed by scanning a large corpus. The method records the frequency of every word co-occurrence with other words in its immediate context (e.g., bi-gram frequency takes 2-word context into account). Each definition is represented by a vector, and the attributes of the vector are the frequencies of the terms that occur in both definitions. The angle between the two vectors is θ. The similarity of the two concepts is defined as cosine(θ). When the relatedness is 1, the two concepts are exactly the same, and when the relatedness is 0, the two vectors are independent of each other without any overlap. Other values in between indicate different degrees of relatedness.

There are two important aspects of the vector method. One is how to construct the definition for the concept. The other is how to find the proper corpus and build the co-occurrence matrix. A comprehensive and accurate definition of each concept is the foundation of our approach. In previous work, Patwardhan et al. [22, 23] used the WordNet-based context vectors to estimate semantic relatedness. This paper presents an extension of this context vector method to the biomedical domain by using the properties of the UMLS. The method includes five steps, 1) count bi-grams, 2) build the co-occurrence matrix, 3) construct concept definitions, 4) calculate semantic relatedness, and 5) evaluate the semantic relatedness using Spearman's rank correlation with human judgments. Figure 2 illustrates the entire procedure. The remainder of this section describes each step in detail.

4.1 Step 1 – bi-grams

The second order context vector measure is a semantic relatedness measure which represents a concept as a context vector. The vector is constructed by counting bi-grams in text within a pre-defined window. For a word *w*, we count the frequency of all two-word pairs (bi-grams) *w*<>*u*. Here, *u* represents words that occur after *w* within the window *l*. The window size dictates how close the bi-gram can occur together. After we scan the entire corpus, the sum of the bi-grams and their frequencies starting with *w* is the first order context vector for word *w*. In this first step, we obtain the bi-grams for every content word in the corpus.

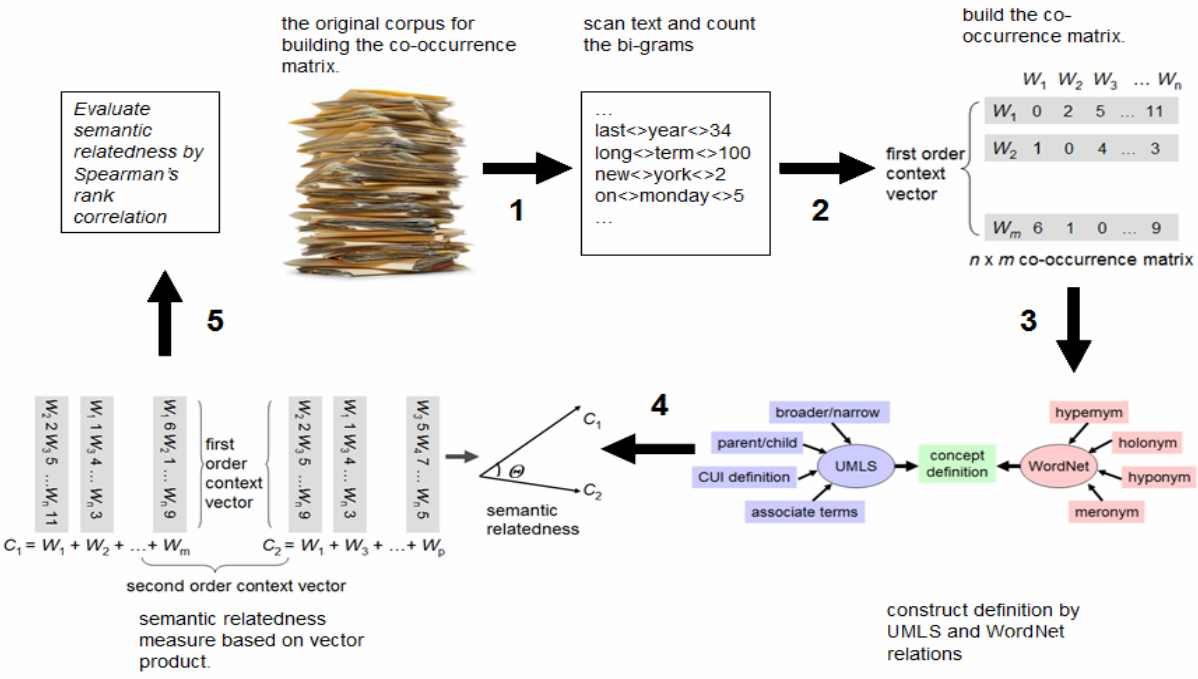


Figure 2. The 5 steps of the second order context vector semantic relatedness method.

4.2 Step 2 – Co-Occurrence Matrix

In the second step, we construct an *m* x *n* co-occurrence matrix which records the frequency of every bi-gram. The matrix is stored in a text file. Each line of the file represents a vector for a word *w*. In order to save the vector more efficiently, we only record the word *w* and its co-occurrence words and their frequencies since most of the cells of the matrix are 0. For example, for w_1, the vector is stored as "w_1 w_2 2 w_3 5 … w_n 11".

4.3 Step 3 – Concept Definitions

We construct the concept definition using the UMLS and WordNet. Concepts in the UMLS are identified by Concept Unique Identifiers (CUIs). However, not all CUIs have adequate definitions. Of the 1,774,202 CUIs in the 2010AB version of the UMLS, only 99,777 have definitions (5.6%). Thus, in addition to the CUI definitions, we experimented with several ways of constructing definitions using relations defined in the UMLS. These relations include parent-child (PAR/CHD) and broader-narrower (RB/RN). Definitions of the associated terms of the CUI (TERM) are also considered. It automatically expands concept

definitions by starting with the CUI's own definition (CUI) and adding to that various combinations of relations. For example, "PAR+CHD" consists of combining all parents and children definitions but not the target concept itself. For the concept *head*, its PAR definition is *anatomical areas of the body;* and its CHD definitions includes *the anterior portion of the head that includes the skin, muscles and structures of the forehead, eyes, nose, mouth, checks and jaw*.

In WordNet, words are represented by a synonym set also called synset. Each synset has an associated definition called a gloss. For example, the gloss of the first sense of the word *hand* is "*hand, manus, mitt, paw – (the (prehensile) extremity of the superior limb; "he had the hands of a surgeon"; "he extended his mitt")*". Synsets are connected to each other through semantic relations such as hypernym, hyponym, meronym and holonym. Banerjee and Pedersen [1] extend the Lesk [15] measure which relies on a synset's definition by also including the definition of its related synsets, referring to it as the extended gloss. We use this extended gloss as WordNet definition. The WordNet definitions are obtained by first extracting all of the CUIs associated terms from the UMLS; second, if any of the terms for a CUI match, the term is considered to be "covered" and the extended gloss of the term's first sense is used as the CUI's definition. The method for constructing these extended definitions constitutes the novel contribution of our approach to the previously developed methods. Definitions of the concepts come from the UMLS and WordNet.

4.4 Step 4 – Semantic Relatedness

The fourth step is to calculate the semantic relatedness between two concepts. The relatedness of two concepts is computed by calculating the cosine of the angle between two vectors [30], as shown in Formula (2):

$$rel_{vector}(c_1, c_2) = \frac{\vec{c}_1 \cdot \vec{c}_2}{(c_1) \cdot (c_2)} \quad (2)$$

c_1 and $\boldsymbol{c}_2$ are two concepts. The context vector is composed of row values in the co-occurrence matrix. Each concept is represented by adding every word's first order context vector. The relatedness of c_1 and c_2 is the cosine of the concept vector c_1 and c_2. Figure 2 step 4 shows a graphical explanation of the method.

4.5 Step 5 – Evaluation

We use Spearman's rank correlation coefficient to assess the relationship between the reference standards and the semantic relatedness results. Spearman's rank correlation, ρ, is a non-parametric (distribution free) measure of statistical dependence between two variables. The assumption is there is no relationship between the two sets of data. The algorithm sorts data in both sets from highest to lowest, and then subtracts the two sets of ranks and gets the difference d. The Spearman's correlation between the ranks is obtained from Formula (3).

$$\rho = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad (3)$$

If there are no repeated data values, a perfect Spearman correlation +1 occurs when each of the variables is a monotone function of the other.

5. EXPERIMENTS

The experiments are developed from four aspects: definition construction, WordNet and the UMLS definition coverage, co-occurrence matrix and corpus selection. These four aspects dominate the experiment results. For the definition construction and coverage, we compare the proposed vector method with the Lesk method. And then, we focus on the vector method to illustrate the influence of the co-occurrence matrix and the corpus.

5.1 Lesk vs. Vector with UMLS Relations

Figure 3 represents the distribution of the Spearman's rank correlation coefficients and definition recalls for different relation combinations of the concept definitions. This test was performed on 300,000 clinical reports (window size 2) without any bi-gram frequency cutoffs.

The recall column represents the percentages of how many pairs out of 430 pairs of concepts have definitions with different UMLS relations. For example, "PAR+CHD" has 346 pairs of concepts with definitions, and the definition recall is 80.5%. Although "TERM" covers 425 pairs of the concepts (recall=98.8%), the relatedness contribution is low (only 0.06 for vector method). Since most CUIs have associated terms ("TERM"), the experiments used this property to combine the associated terms' definitions from WordNet in section 5.2. Because of high definition recall and Spearman's correlation, we choose the relation combination to construct the definition.

In Figure 3, the vector and Lesk [15] columns show the Spearman's correlation. We could observe the similar pattern of these two ontology-independent methods. However, after optimizing the low and high bi-gram frequency cutoffs, the vector method obtained higher correlations to human judgments than the Lesk method in the following experiments.

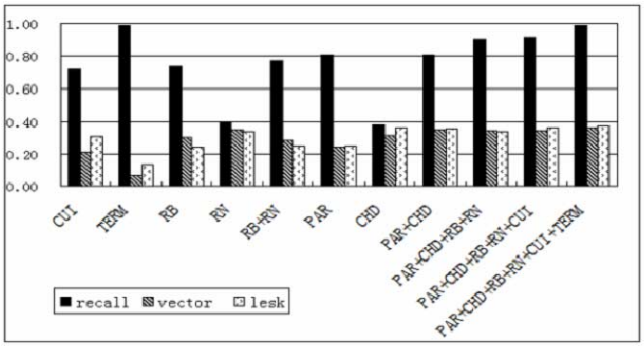


Figure 3. Spearman's rank correlation and definition recall with different relationship combinations of the concept definitions for the vector and Lesk method.

5.2 Lesk vs. Vector with WordNet and UMLS Coverage

The general English terminological system WordNet and the domain specific UMLS have different coverage of the medical concepts. For the 351 unique CUIs in the 430 pairs of concepts, 294 CUIs (83.76%) have a UMLS definition and 284 (80.91%) have a definition in WordNet. Combining them the coverage is 322 out of 351 (91.73%).

Table 2 shows the coverage of the UMLS and WordNet relative to the "CUI" and "PAR+CHD+RB+RN+CUI" relation in the UMLS. 'WordNet' means the definitions come from CUIs' mapping

definition in WordNet. 'UMLS' means the definitions came from the UMLS definition via different relations. 'Combination' means both WordNet and the UMLS were used. For total of 430 pairs of concepts, 276 pairs of CUIs have definitions in WordNet. 307 pairs have definitions in the UMLS CUIs and 393 pairs have definitions in the "PAR+CHD+RB+RN+CUI" relationship. Clearly, the UMLS has some definitions that WordNet doesn't have, and vice versa, as well as some concepts defined in both. The improvement in coverage becomes fairly dramatic with added relations. The coverage for CUI is 379 pairs and 406 pairs (94.4%) for the extended relation. We also proved the combination of the UMLS and WordNet had major impact on the performance of the vector method in next section 5.6.

Table 2. UMLS and WordNet definition coverage

	CUI	PAR+CHD+RB+RN+CUI
WordNet	276	276
UMLS	307	393
Combination	379	406

Table 3 shows the results of experiments with this new approach for constructing definitions. For the Lesk method, using WordNet obtains a higher correlation than using the extended definitions of the UMLS (0.365 vs. 0.232. Since the definitions are only from WordNet, the relatedness scores are independent from the relations.) For the vector method, however, the opposite is true. Relatedness scores are very close for the "CUI" definition (0.309 vs. 0.288). When the relationship is extended to "PAR+CHD+RB+RN+CUI", Lesk performs worse than vector (0.362 vs. 0.421). Adding WordNet definitions to the UMLS definitions increases the correlation results for vector (0.421 vs. 0.461).

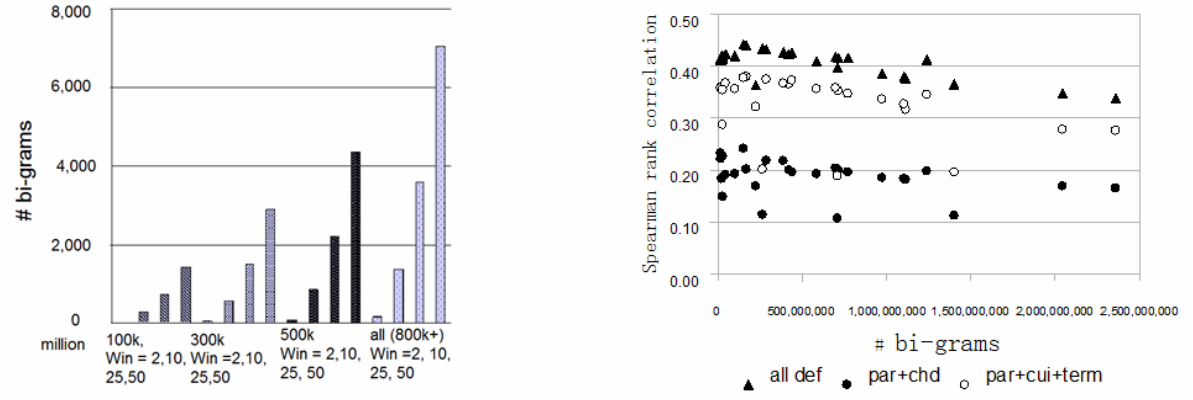
The results indicate that the choice of the method for extending concept definitions has the greatest effect of the performance of the vector approach. Using a number of semantic relations in the UMLS in addition to the WordNet definitions works the best for the second order co-occurrence vector method.

Table 3. The UMLS and WordNet semantic relatedness results

	CUI		PAR+CHD+RB+RN+CUI	
Method	**Lesk**	**vector**	**Lesk**	**vector**
WordNet	0.365	0.232	0.365	0.232
UMLS	0.309	0.288	0.362	0.42
Combination	0.384	0.294	0.404	**0.46**

5.3 Vector with bi-grams

The number of bi-grams is directly related to three parameters: (1) corpus size, (2) window size, and (3) cut-off threshold for removal of low and high frequency bi-grams. We extract the bi-grams with different document sizes and window sizes. Figure 4a illustrates the increase in bi-grams with different documents and windows. We can clearly see that the number of bi-grams increases exponentially with the increase of the window size. Figure 4b represents the distribution of the Spearman's rank correlation for three different definitions. The correlation first increases and then reaches the maximum around 100 million bi-grams. Beyond that, the correlation begins to decrease. This pattern is observed for all three different types of definitions.



(a) (b)

Figure 4. (a) Number of bi-grams. (b) Spearman's rank correlation.

The results in Figure 4 show that the "all" method achieves the highest correlation with the reference standard; however, for the three extended definitions, we can observe that the correlation scores were related to the total number of bi-grams used to construct the co-occurrence matrix. The number of bi-grams is controlled by the corpus size, window size, and two frequency thresholds – low and high. From Figure 3, we observe that the greater the number of bi-grams, the lower the correlation. Large text with lower window sizes could generate the same amount of bi-grams. Usually large text plus lower window sizes result in a better correlation. The experiments use window 2 because it counts the most adjacent bi-grams frequency for every word. Low frequency bi-grams only appear in few clinical reports. Thus, when we construct the vectors, they do not contribute sufficient information. High frequency bi-grams appear in many clinical reports. They do not differentiate one vector from another, but

high frequency bi-grams can emphasize the relatedness of two concepts. In Table 4, we list the percentages with different low/high frequency cutoffs. The relationship between the window size and the low frequency should be considered also. When using window size 10, most bi-grams will occur at least 10 times. So, removing bi-grams with low frequency of 10 or less is very conservative. Using greater low frequency thresholds result in better results.

The tf-idf weight is another statistical measure for evaluating how important a word is to a document. The reason we do not use tf-idf is because tf-idf could remove potentially important words completely but bi-gram cutoffs only remove low/high frequency bi-grams of a word, and the word itself won't be removed.

Table 4. Percentages of bi-grams retained after applying thresholds for 100k clinical reports with window size 10

low freq. cutoffs	freq<2	freq<10	freq<15
%bi-grams	94.34%	81.91%	78.26%
high freq. cutoffs	freq>500	freq>800	freq>1000
%bi-grams	64.29%	69.81%	72.29%

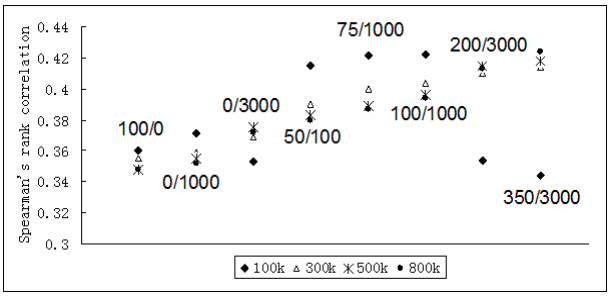


Figure 5. Vector with Spearman's rank correlation for different document sizes and bi-gram low/high frequency cutoffs.

The Spearman's rank correlation was also related to the total number of bi-grams used to construct the co-occurrence matrix. Figure 5 shows the correlation changes with the size of MEDLINE documents and different bi-gram cutoffs. We then used the combination of "PAR+CHD+RB+RN+CUI" to build concept definitions. For example, after removing bi-grams with frequencies less than 100 and more than 1000 from a corpus of 100k documents, we achieved a correlation score of 0.422 (Further bi-gram frequency cutoffs for 300k, 500k and 800k documents did not improve the results.)

5.4 Vector with Clinical Reports vs. MEDLINE Abstracts

When the experiments use MEDLINE (300k documents) abstracts to build the co-occurrence matrix, we observe similar experimental results as when using the clinical reports except with TERM and RB. The advantage of using MEDLINE over clinical reports to calculate semantic relatedness is that MEDLINE is freely available. (Figure 6)

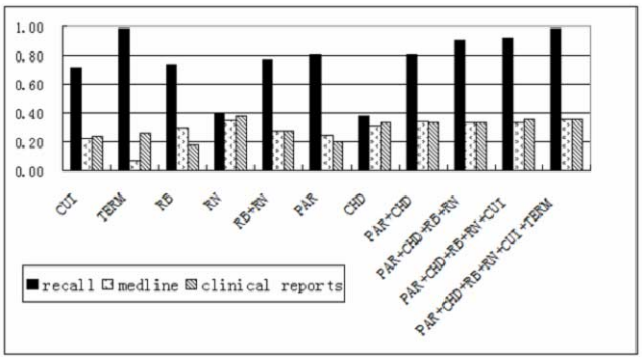


Figure 6. Vector with Clinical reports vs. MEDLINE abstracts.

When the corpus was changed to English Gigaword, the result of the vector-based relatedness score dropped from 0.46 to 0.25. This result suggested that effective computation of semantic relatedness for medical concepts requires corresponding medical corpora to build the co-occurrence matrix.

6. DISCUSSION

The focus of the vector method is the concept construction. This is based on the assumption that one definition might not fully represent the meaning of the concept. The combination of general English and domain specific English resources, such as WordNet and the UMLS, could make up for the insufficient information in individual resources. Every high dimensional vectors of an individual concept represents its distributional semantics in a larger corpus. This flexible format makes the vector method free from the ontology structure and it is also easy to adapt to a new domain.

The experiments use Spearman's rank correlation to compare the relative performance of the various methods to human judgments. Humans use the dominant sense of the target words or the related senses triggered mutually to compare how the concepts are related [5]. Computer methods, either ontology-dependent or ontology-independent measures, exhibit that concepts can be related in many different ways. Human raters observed one sort of relatedness, while the proposed vector measure potentially finds another kind of relatedness. Lack of significant correlation does not necessarily mean that the computer methods are not doing something useful - just that whatever they are doing is different from human judgments.

Table 5 gives the three top and bottom pairs of concepts ranked by the semantic relatedness. The usefulness of semantic relatedness methods also needs to be evaluated using indirect validation approaches such as information retrieval and word sense disambiguation. An acronym word sense disambiguation project showed the usefulness of the proposed vector method. It achieved an overall accuracy of 89% [19].

When comparing the vector relatedness results with the path similarity measure on the same dataset, the "path" measure [28] yields a much lower correlation (r=0.29) than the vector-based method (r=0.46). Other path based methods, proposed by Wu & Palmer [32] (r=0.24) and Leacock & Chodorow [14] (r=0.29), also have a lower correlation than the vector method. This is due to the fact that the path-based approach relies exclusively on hierarchical relations.

Table 5. Semantic relatedness of 3 top and bottom pairs of concepts

Relatedness	Top 3 pairs
1	ferrous gluconate<>ferrous fluconate
0.9932	ceftazidime<>ceftriaxon
0.9918	cefaclor<>cefoxitin
Relatedness	**Bottom 3 pairs**
0.0152	scapulalgia<>nitroglycerine
0.0126	cefalea unilateral (hallazgo)<>nitroglycerol
0.0072	waterbrash<> regurgita luego de deglutir

In the future, we plan to use the semantic relatedness measures to evaluate semantic labels created by Semantic Knowledge Representation (SemRep) of National Library of Medicine. Semantic relatedness is going to help us group similar relationships. The relationship discovered from electronic medical records and electronic therapeutic records will also help us to find adverse drug events.

7. SOFTWARE RESOURCES

The software for the vector measure is part of UMLS-Similarity which is an open source software package [18] and can be download from CPAN[4]. It consists of a suite of Perl modules that can be used to calculate the similarity/relatedness between two concepts based on the structure and content of the UMLS. It provides a command line interface, API, and web interface. Some of the measures in this package were originally developed for WordNet and are implemented in the WordNet-Similarity package [25]. The WordNet-Similarity package inspired the creation of the UMLS-Similarity package but the structure and nature of the UMLS is completely different from WordNet, and the adaptation of those measures was not straightforward. The core backbone of the package is completely different and offers specific functionality to the UMLS but not available in WordNet.

The web interface[5] provides a subset of the functionality offered by the API and command line. The purpose of the web interface is to demonstrate the functionality of UMLS-Similarity without requiring the user to install the UMLS in a MySQL database. It provides a simple way to introduce the package's source and relation.

8. CONCLUSIONS

This paper describes a procedure to construct the second order co-occurrence vector measure for semantic relatedness. We discuss the influences of corpus selection, size of the corpus, low and high bi-gram cutoffs, and the concept definition construction. The method for extending concept definitions has the greatest effect of the performance. The paper also exhibits the different coverage of concept definitions between the UMLS and WordNet. Using the proper relations in the UMLS plus WordNet definitions can obtain a high correlation with human judgements.

[4] CPAN: www.cpan.org

[5] UMLS-Similarity web interface: http://atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi

9. ACKNOWLEDGMENTS

This work was supported by the Grant #R01 LM009623-01 from the National Library of Medicine. We would like to thank Fairview Health Services for the use of their clinical data.

10. REFERENCES

[1] Banerjee, S. and Pedersen, T. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,* Acapulco, Mexico, 805-810.

[2] Bodenreider, O. and Burgun, A. 2002. Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In *Proceedings of the First Global WordNet Conference*, 223–230.

[3] Bodenreider, O. 2004. *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Research, 32, D267-D270.

[4] Bousequet, C., Lagier, G., LilloLe, L.A., Le Beller, C., Venot, A., and Jaulent, M.C. 2005. Appraisal of the MedDRA Conseputal Structure for describing and grouping adverse drug reactions. *Drug Safety*, 28(1), 19-34.

[5] Budanitsky, A. and Hirst, G. 2006. Evaluation WordNet-based measures of lexical semantic relatedness. *Journal of computational linguistics*, 32(1), 13-47.

[6] Burgun, A. and Bodenreider, O. 2001. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In *Proceedings of NAACL Workshop*, 77-82.

[7] Chen, B., Foster, G., and Kuhn, R. 2010. Bilingual sense similarity for statistical machine translation. In *Proceedings of the ACL*, 834-843.

[8] Erk, K. 2007. A simple, simple-based model for selectional preferences. In *Proceedings of the ACL*, 858-865.

[9] Guo, X., Liu, R., Shriver, C.D., Hu, H., and Liebman, M.N. 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Journal of Bioinformatics,* 22(8), 967-973.

[10] Hirst, G. and St-Onge, D. 1998. *Lexical chains as representations of context for the detection and correction of malapropisms.* In WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, 305-332.

[11] Hogan, D. 2007. Empirical measurements of lexical similarity in noun phrase conjuncts. In *Proceedings of the ACL*. 680-687.

[12] Huang, K.C., Geller, J., Halper, M., Perl, Y., and Xu, J. 2009. Using WordNet synonym substitution to enhance UMLS source integration. *Journal of Artificial Intelligence in Medicine*, 46(2):97-109.

[13] Jiang, J. and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, 19-33.

[14] Leacock, C. and Chodorow, M. 1998. *Combining local context and WordNet similarity for word sense identification.* In WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, 265–283.

[15] Lesk, M. 1986. 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*, Toronto, Canada, 24-26.

[16] Lin, D. An information-theoretic definition of similarity. 1998. In *Proceedings of the International Conference on Machine Learning,* 296-304.

[17] Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. 2003. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Journal of Bioinformatics,* 19(10),1275-1283.

[18] McInnes, B, Pedersen T, Pakhomov S. UMLSInterface and UMLS-Similarity: Open Source Software for measuring paths and semantic similarity. In *Proceedings of AMIA*, 431-435.

[19] McInnes, B., Pedersen, T., Liu, Y., Pakhomov, S., and Melton, G. 2011. Using Second-order vectors in a Knowledge-based Method for Acronym Disambiguation. In *the Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 145-153.

[20] Mougin F., Burgun, A., and Bodenreider, O. 2006. Using WordNet to Improve the Mapping of Data Elements to UMLS for Data Sources Integration. In *Proceedings of AMIA*, 574-578.

[21] Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *Proceedings of AMIA*, 572-576.

[22] Patwardhan, S. and Pedersen, T. 2006. 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, 1-8.

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

[24] Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, C. 2006. Measures of Semantic Similarity and Relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3), 288-299.

[25] Pedersen, T., Patwardhan, S., and Michelizzi, J. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT NAACL*, 38-41.

[26] Ponzetto, S. and Strube, M. 2007. An API for measuring the relatedness of words in Wikipedia. In *Proceedings of the ACL*, 49-52.

[27] Pucher, M. 2007. WordNet-based semantic relatedness measures in automatic speech recognition for meetings. In *Proceedings of the ACL*, 129-132.

[28] Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics,* 19(1):17-30.

[29] Resnik, P. 1995. Using information content to evaluate semantic similarity. I*n Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 448-453.

[30] Schutze, H. 1998. Automatic word sense discrimination. *Compute Linguist*, 24(1), 97-124.

[31] Tsang, V. and Stevenson, S. 2010. A graph-theoretic framework for semantic distance. *Journal of Computational Linguistics*, 36(1), 31-69.

[32] Wu, Z. and Palmer, M. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Meeting of Association of Computational Linguistics*, Las Cruces, NM, 133-138.