Semantic Similarity Using First and Second Order Co-occurrence Matrices and Information Content Vectors

AHMAD PESARANGHADER
Faculty of Creative Multimedia
Multimedia University
Jalan Multimedia, 63100 Cyberjaya
MALAYSIA
ahmad.pgh@gmail.com

SARAVANAN MUTHAIYAH
Faculty of Management
Multimedia University
Jalan Multimedia, 63100 Cyberjaya
MALAYSIA
saravanan.muthaiyah@mmu.edu.my

Abstract: - Massiveness of data on the Web demands automated Knowledge Engineering techniques enabling machines to achieve integrated definition of all available data to make a unique understanding of all discrete data sources. This research deals with Measures of Semantic Similarity resolving foregoing issue. These measures are widely used in ontology alignment, information retrieval and natural language processing. The study also introduces new normalized functions based on first and second order context and information content vectors of concepts in a corpus. By applying these measures to Unified Medical Language System (UMLS) using WordNet as a general taxonomy and MEDLINE abstract as the corpus to extract information content and information content vectors, these functions get evaluated against a created test bed of 301 biomedical concept pairs scored by medical residents. The paper shows newly proposed Semantic Similarity Measures outperform previous functions.

Key-Words: - Semantic Similarity, Computational Linguistic, UMLS, WordNet

1 Introduction

Measures of semantic similarity and relatedness functions aim at finding if one pair of concepts (or documents) is more related than another considering human ability for this judgement. These functions have a wide usage in ontology matching [1], machine translation [2], automatic speech recognition [3], and text categorizing. They can also be effectively applied in semantic searching of textual resources available for both general and specific domains of knowledge. The output of a similarity measure is a value, ideally normalized between 0 and 1 inclusive, indicating how much two given words (or documents) are semantically similar. Richness of document corpora, used for information extraction (e.g. bi-grams) feeding to functions, would enhance similarity function performance; therefore, meticulousness in corpus selection for a specific application is essential.

This paper presents existing Measures of Semantic Similarity already proposed. The study also introduces new normalized functions for measuring Semantic Similarity based on first and second order context and information content vectors of concepts in a corpus. The intuition behind these new methods named *First* and *Second Order Context Vector Similarity Measures* is that terms surrounding a concept in a context often carry the same sense of that concept. For example, *skull* is more similar to *cranium* than to *retina* because *skull* and *cranium* can share the same surrounding context. The aim of the study is to determine whether considering the context of the concept instead of the concept itself in a corpus leads to a more effective result.

To achieve research objectives, it is shown how already proposed semantic similarity functions and our proposed measures can be adapted and compared in the biomedical domain. These measures are implemented by considering MeSH ontology included in UMLS Metathesaurus, taking advantage of WordNet as a general taxonomy and MEDLINE abstract as the corpus for extraction of information content and information content vectors. These measures are evaluated against a created test bed of 301 biomedical concept pairs already scored by eight medical residents. We will conclude that newly proposed semantic similarity measures would outperform other measures specifically *Lin* measure as the baseline in the study with respect to the Spearman's rank correlation coefficient results in any case.

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 similar terms automatically by using information from a large corpus and existing ontologies. This study coping with semantic similarity issue considers it as a distinct and specific case of semantic relatedness [4].

Existing semantic similarity measures can be categorized into three groups:

2.1 Path Based Semantic Similarity Measures

In path based measures distance between concepts (nodes) on the hierarchy or thesauruses are intuitively appealing. In other words the only factor important for this measure type is the shortest number of jumps from one concept to another concept.

The proposed semantic similarity measures based on this approach are:

- Rada et al., 1989 (*path*) [5]

$$\boldsymbol{sim_{path}}(\mathbf{c_1}, \mathbf{c_2}) = \frac{\mathbf{1}}{\mathbf{shortest\ is-a\ path}(\mathbf{c_1}, \mathbf{c_2})} \quad (1)$$

- Caviedes & Cimino, 2004 (*cdist*) [6]

In path-based measures we count nodes (not paths). This formula is normalized as the minimum and maximum possibility of similarity is 0 and 1 respectively.

2.2 Path Based and Depth Based Semantic Similarity Measures

Path-based methods doesn't account for specificity whereas deeper as well as informative paths tend to travel less semantic distance. To overcome this drawback of path-based methods new methods based on both path and dept of concepts are proposed. These measures are:

- Wu & Palmer, 1994 (*wup*) [7]

$$\boldsymbol{sim_{wup}}(\mathbf{c_1}, \mathbf{c_2}) = -\mathbf{log}\left(\frac{\mathbf{2 \times depth}(\mathbf{LCS}(\mathbf{c_1}, \mathbf{c_2}))}{\mathbf{depth}(\mathbf{c_1}) + \mathbf{depth}(\mathbf{c_2})}\right) \quad (2)$$

(LCS is the least common subsumer of the two concepts)

- Leacock & Chodorow, 1998 (*lch*) [8]

$$\boldsymbol{sim_{lch}}(\mathbf{c_1}, \mathbf{c_2}) = -\mathbf{log}\left(\frac{\mathbf{minpath}(\mathbf{c_1}, \mathbf{c_2})}{\mathbf{2D}}\right) \quad (3)$$

(*minpath* means shortest path, and *D* is the total depth of the taxonomy)

- Zhong et al., 2002 (*zhong*) [9]

$$\boldsymbol{dist_{zhong}}(\mathbf{c_1}, \mathbf{c_2}) = \frac{\mathbf{2 \times m}(\mathbf{LCS}(\mathbf{c_1}, \mathbf{c_2}))}{\mathbf{m}(\mathbf{c_1}) + \mathbf{m}(\mathbf{c_2})} \quad (4)$$

$$(\ \boldsymbol{m}(\mathbf{c}) = \frac{\mathbf{1}}{\mathbf{k}^{(\mathbf{depth}(\mathbf{c})+\mathbf{1})}}\ )$$

- Nguyen & Al-Mubaid, 2006 (*nam*) [10]

$$\boldsymbol{sim_{nam}}(\mathbf{c_1}, \mathbf{c_2}) = -\mathbf{log}(\ \mathbf{2} + (\mathbf{minpath}(\mathbf{c_1}, \mathbf{c_2}) - \mathbf{1}) \times (\mathbf{D} - \mathbf{d}))\ ) \quad (5)$$

(*D* is the total depth of the taxonomy, and $\boldsymbol{d} = \mathbf{depth}(\mathbf{LCS}(\mathbf{c_1}, \mathbf{c_2}))$)

2.3 Path Based and Information Content Based Semantic Similarity Measures

Depth shows specificity but not frequency, meaning that low frequency concepts often are much more specific and informative than high frequency ones. This quantity for one concept technically is generally known as Information Content [11]. IC is formally defined as the negative log of the probability of a concept on the ontology calculable using an external corpus.

$$\boldsymbol{IC}(\mathbf{c}) = -\mathbf{log}(\ \boldsymbol{p}(\boldsymbol{c})\ ) \quad (6)$$

(*p(c)* is probability of a concept)

$$\boldsymbol{p}(\mathbf{c}) = \frac{\mathbf{tf} + \mathbf{if}}{\mathbf{N}} \quad (7)$$

(*tf* is term frequency or frequency of concept itself, *if* is inherited frequency or frequency of concept's descendants in total, and *N* is sum of all concepts' frequencies on the ontology)

Methods based on IC of concepts are:

- Resnik, 1995 (*res*) [11]

$$\boldsymbol{sim_{res}}(\mathbf{c_1}, \mathbf{c_2}) = \mathbf{IC}(\mathbf{LCS}(\mathbf{c_1}, \mathbf{c_2})) \quad (8)$$

- Jiang & Conrath, 1997 (*jcn*) [12]

$$sim_{jcn}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2))} \quad (9)$$

- Lin, 1998 (*lin*) [13]

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (10)$$

For measures calculated based on information content, *res* and *jcn* are non-normalized while *lin* is normalized.

In this research all foregoing measures are tested and evaluated on existing reference standard. The MEDLINE is used as a corpus for populating appropriate information content file. However, central part of the study constitutes implementation and examination of new similarity measures. As previously stated, the main idea behind these new methods, called *First* and *Second Order Context Vector Similarity Measures* is that terms surrounding a concept in a context often convey the same sense of that concept. These proposed measures in some way are an extended and combined version of the *lin* idea for semantic similarity measurement and *vector* (gloss vector) idea [14] for semantic relatedness calculation.

3 Experimental Data

Some of external resources available for the study act as thesauruses; other resources are a corpus to extract required information feeding to the semantic similarity measure, and a dataset used for testing. While MEDLINE abstract is used as a corpus, the UMLS and WordNet are used for construction of definitions for concepts. The resources briefly are:

3.1 Unified Medical Language System (UMLS)

The Unified Medical Language System[1] (UMLS) is a knowledge representation framework designed to support biomedical and clinical research. Its fundamental usage is provision of a database of biomedical terminologies for encoding information contained in electronic medical records and medical decision support. It comprises over 161 terminologies and classification systems. The UMLS contains more than 2.6 million concepts and 8.6 million unique concept names. The three foremost components of the UMLS are the Metathesaurus, Semantic Network and SPECIALIST Lexicon.

Basically this research focuses on the Metathesaurus since for calculation of all semantic similarity methods examined in the study we need to have access to biomedical concepts resided on UMLS Metathesaurus. Some of terminologies (sources) contained in the UMLS include National Cancer Institute Thesaurus (NCIT), SNOMED Clinical Terms (SNOMED CT), and Medical Subject Headings (MeSH). 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. While all concept pairs tested in this study are from MeSH, the accessibility to MeSH on UMLS is requisite for our experiments. Basically MeSH is a comprehensive controlled vocabulary aiming at indexing journal articles and books in the life sciences; additionally, it can serve as a thesaurus that facilitates searching. In this research we limited the scope to 2011AB release of the UMLS.

3.2 WordNet

WordNet[2] is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

Most synonym sets are connected to other synsets via a number of semantic relations. These relations vary based on the type of word, and include: hypernym, hyponym, coordinate terms, holonym and meronym for nouns; hypernym, troponym, entailment and coordinate terms for verbs; related nouns, similar to and similar to for adjectives; and root adjectives for adverbs.

WordNet provides a certain number of medical terms; it is shown that the concept overlap between WordNet and the UMLS changes from 48% to 97%

[1] http://www.nlm.nih.gov/research/umls

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

[15]. The reason is that WordNet only records the canonical forms of vocabularies, while the UMLS records the variability of the lexical forms encountered in the source vocabularies. The study makes used of WordNet 3.1.

3.3 Text Corpus - MEDLINE Abstract

MEDLINE[3] is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution. The database contains more than 21.6 million records from 5,582 selected publications covering biomedicine and health from 1950 to the present. It uses Medical Subject Headings (MeSH) for information retrieval.

For the current study we used MEDLINE article abstracts as the corpus to build a term-term co-occurrences matrix for subsequent computation of semantic similarity. We used the 2012 MEDLINE abstract. Table 1 demonstrates number of extracted single-words and bi-grams required as supplementary information for semantic similarity measures' calculation.

Table 1
The Comparison of MEDLINE Bi-grams and Single Words

	Including Stop Words	**Excluding Stop Words**	**Size (MB) Including SW / Excluding SW**
Single Words	1184909396	1124722499	36 MB / 34 MB
Bi-grams	547316681	479567529	962 MB / 826 MB

3.4 Reference Standard

The reference standard[4] used in our experiments was based upon a set of medical pairs of terms created specifically for testing automated measures of semantic similarity freely provided by University of Minnesota Medical School as an experimental study [16]. In their study 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. Four medical residents (2 women and 6 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 similarity 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 - 0.47); however, in order to reach a good agreement, after removing some concept pairs, a set of 566 UMLS concept pairs manually rated for semantic similarity using a continuous response scale was provided.

Some of concepts from original reference standard are not included into the MeSH Ontology. Therefore, after removing those concepts from this dataset, a subset of 301 concept pairs for testing on different semantic similarity functions including new measures in this study was available.

4 Methods

Proposed measures in this study are named *First Order Context Vector Similarity Measure* and *Second Order Context Vector Similarity Measure*. These methods are dependent on first order co-occurrence matrix and second order co-occurrence matrix created form corpus (MEDLINE) consecutively. In order to build these matrices we would consider and record 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). For building second order co-occurrence matrix the definition of a concept (each row of the matrix) is represented by a vector, calculated through summation of all first order vectors of constituent words in the definition. The first and second order co-occurrence matrices would help for population of information content matrices used in our similarity measures. The method for constructing an information content matrix constitutes the novel contribution of our approach to the previously developed methods. Considering measure of relatedness for two concepts having each concept's definition, the basic idea is comparing these two definitions (angle of two vectors on the second order co-occurrence matrix) [17]. We would borrow this idea for *Second Order Context Vector Similarity Measure* (instead of relatedness) in such a way that the information content of two concepts' definitions together would

[3] http://mbr.nlm.nih.gov/Download/index.shtml

[4] http://rxinformatics.umn.edu/data/UMNSRS_similarity.csv

be compared with their least common subsumer (LCS) definition's information content. In other words, the similarity of the two concepts can be defined as cosine(θ) between their information content united together and information content of their LCS. When the similarity is 1, the two concepts are exactly the same, and when the similarity is 0, they are strongly unlike. Other values in between indicate different degrees of similarity.

For *First Order Context Vector Similarity Measure* the procedure is the same with a slight difference that we would use directly first order co-occurrence matrix for calculation of information content which means no need for definition construction phase there. Be aware that the information content for two proposed measures in this study would be in the form of a vector instead of a scalar that we used to have in previous methods. The information content vector for a concept is (element-wise) negative log of probably vector of that concept. This probability vector is computable having concept's vector from first or second order co-occurrence matrix augmented with vectors of the descendants of that concept on the ontology divided by sum of all concepts' vectors on the ontology.

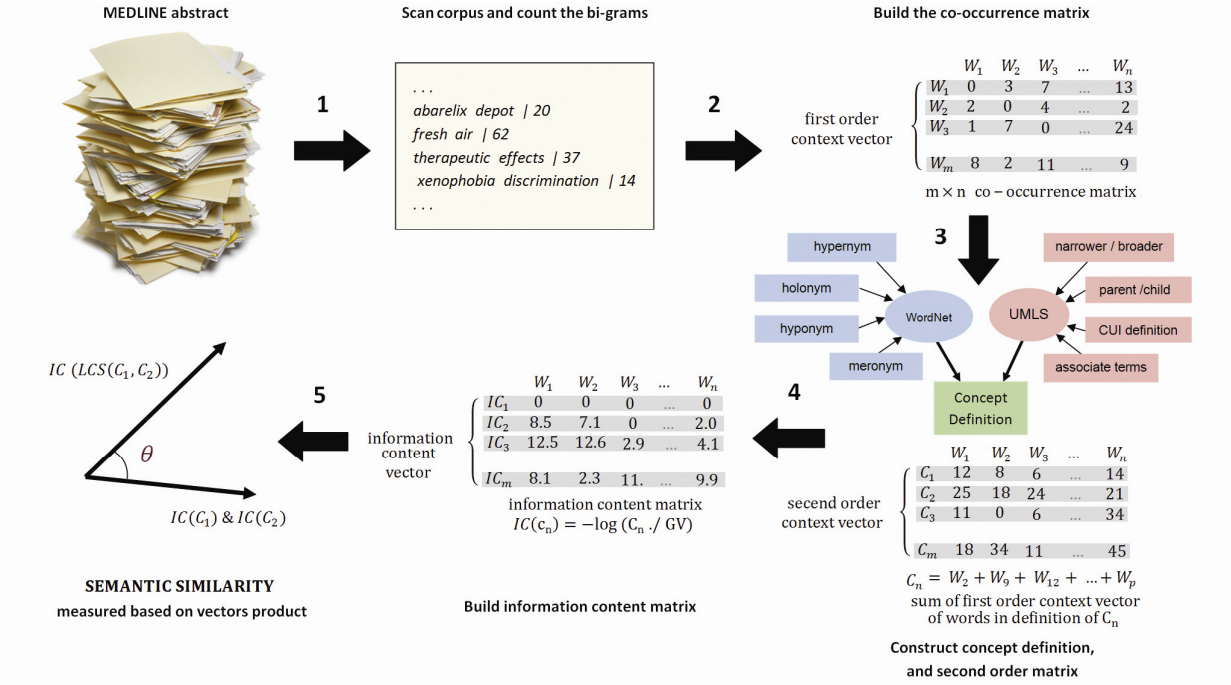

Fig.1 The 5 steps of the First and Second Order Context Vector Semantic Similarity Methods

There are three important aspects of our methods. One is how to construct the definition for the concept which is important in *Second Order Context Vector Similarity*. The second is how to find the proper corpus and build the co-occurrence matrix, and the third how to construct information content matrix. Totally our proposed method includes five steps, 1) count bi-grams, 2) build the first order co-occurrence matrix, 3) construct concept definitions and second order co-occurrence matrix, 4) build the information content matrix, and 5) calculate semantic similarity. For the *First Order Context Vector Similarity Measure* we would have third phase eliminated, therefore in the fourth phase we would build the information content matrix immediately from first co-occurrence matrix (instead of second co-occurrence matrix). Figure 1 illustrates the entire procedure. The remainder of this section describes each step in detail.

4.1 Step 1 – bi-grams

The *First/Second Order Context Vector Similarity Measure* is a semantic similarity measure which represents a concept as a context vector. The vector is build trough counting bi-grams in text within a window. For a word *w*, we tally the occurrence of all two-word pairs (bi-grams) *w u*. Here, *u* indicates words that appear right after *w* within the window *2*. The window size controls how close two words can

appear in bi-grams. In first step, after scanning the corpus entirely, all bi-grams and their frequencies for all content words would be calculated.

4.2 Step 2 – First Order Co-Occurrence Matrix

In this step, we construct first order co-occurrence matrix by having frequency of all bi-grams for each word available from first step. The matrix is stored in a text file while each line of the file represents a vector for a word *w*. Since the matrix is sparse (most of the cells of the matrix are 0), we only record the word *w,* its co-occurrence words and their frequencies. For example, for w_1, the vector is stored as " w_1 w_2 2 w_4 7 w_8 1 … w_{28} 3 ".

4.3 Step 3 – Concept Definitions and Second Order Co-occurrence Matrix

The step is for *Second Order Context Vector Similarity Measure* only, in which 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. Thus, in addition to the CUI definitions, according to second order context vector relatedness measure [17] we would examine 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.

In WordNet, words are characterized by a synonym set also called synset which has its own associated definition named a gloss. Synsets are connected to each other through semantic relations such as hypernym, hyponym, meronym and holonym. Banerjee and Pedersen [18] for the first time extend the Lesk [19] 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. For WordNet gloss of a concept appropriate sense pertaining to that concept in WordNet was selected.

After having all available definitions for concepts in a targeted source of UMLS Metathesaurus, it would be possible to construct second order co-occurrence matrix derived from these definitions. The procedure briefly is that for each word in the definition, we already have first order co-occurrence vector based on the occurrence of that word with other words in an external corpus (MEDLINE). For second order co-occurrence vector of a term (concept) we take the centroid of the vector associated with each word in the definition of that term. For the *First Order Context Vector Similarity Measure* we would ignore this step and use the first order co-occurrence vector of term directly in order to construct information content matrix.

4.4 Step 4 – Information Content Matrix

Here we would construct information content matrix. In previous measures relying on information content we had this value in scalar (numeric) form. In our proposed measures we would have this value represented as a vector. The collection of these vectors would build information content matrix. The formula to construct the information content matrix is:

$$\boldsymbol{IC_{matrix}}(\mathbf{c}) = -\mathbf{log}\,(\mathbf{CV}(\mathbf{c})\,./\,\mathbf{GCV}) \qquad (11)$$

Where CV is first/second order context vector of the concept *c*, available in first/second order co-occurrence matrix, augmented by firsts order context vector of its subordinates (its descendants). GCV is global first order context vector reachable through summation of all first/second order context vectors of words in the targeted source in the UMLS. Notation ./ mathematically indicates "dot division" meaning division of two matrices element-wise.

4.5 Step 5 – Semantic Similarity

The fifth step is to calculate the semantic similarity between two concepts. The similarity of two concepts is computed by calculating the cosine of the angle between two vectors; first vector would be united form of two concepts information content vectors and second vector would be equivalent information content vector of be the least common subsumer (LCS) of two concepts. The formula for similarity calculation would be:

$$\boldsymbol{sim_{vector}}(\mathbf{c_1}, \mathbf{c_2}) =$$

$$\mathbf{cosin}\left(\mathbf{IC}_{\boldsymbol{vec}}\left(\mathbf{LCS}(\mathbf{c_1}, \mathbf{c_2})\right), \mathbf{sqrt}\left(\mathbf{IC}_{\boldsymbol{vec}}(\mathbf{c_1}) .\times \mathbf{IC}_{\boldsymbol{vec}}(\mathbf{c_2})\right)\right) \qquad (12)$$

In the formula c_1 and c_2 indicate two input concepts. The IC_{vec} is equivalent to information content vector for each concept. LCS is least common subsumer for two concepts on the ontology (MeSH). The sign indicates "dot multiplication" (multiplication of two matrices element-wise) and *sqrt* function returns the square root of each element of a vector.

5 Metric

In this project Spearman's rank correlation coefficient to assess the relationship between the reference standards and the semantic similarity results will be applied. Spearman's rank correlation, , is a non-parametric (distribution free) measure of statistical dependence between two variables. Here we assume that there is no relationship between the two sets of data. This 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 attainable through formula:

If there are no repeated data values, an exact Spearman correlation +1 occurs which means each of the variables is a monotone function of the other.

6 Experiments

The experiments are developed from three aspects: definition construction, bi-grams size, and co-occurrence matrix. These three aspects dominate the experiment results. After the definition construction used for *Second Order Context Vector Similarity Measure*, we compare the proposed similarity methods with the other similarity methods already presented, then, we focus on the *First/Second Order Context Vector Similarity Measure* to illustrate the influence of the co-occurrence matrix populated from external resource.

6.1 First/Second Order Similarity vs. Others Measures

Figure 2 represents the distribution of the Spearman's rank correlation coefficients on 301 concept pairs for similarity measures proposed already and in this research.

The test is done on MEDLINE abstract (window size 2) including removal of stop words and without any bi-grams frequency cut-offs. The features and formula for all similarity measures on figure are represented in the paper earlier.

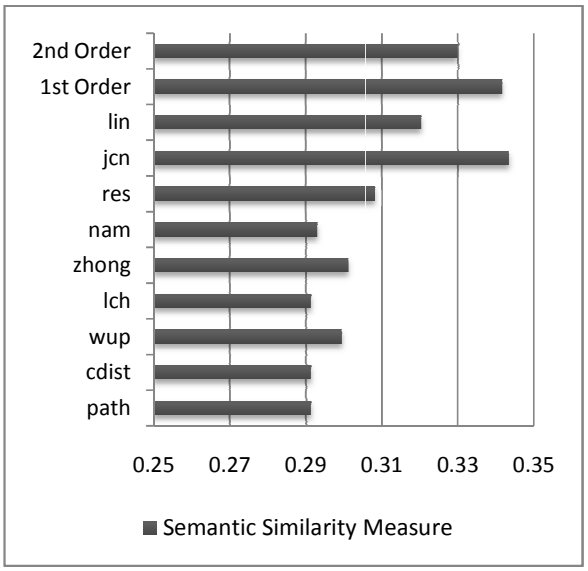

Fig.2 Spearman's correlation for semantic similarity measures

Table 2 represents precisely the Spearman's rank correlation for 5 highest semantic similarity measure results.

Table 2
Semantic Similarity Correlation Results of 5 Highest Methods

Semantic Similarity	Spearman's Rank Correlation
res measure	0.3082
jcn measure	0.3434
lin measure	**0.3203**
1st Order Similarity Measure	**0.3416**
2nd Order Similarity Measure	0.3299

The similarity method *jcn* has the highest correlation between calculated results of concept pairs and human judgment of them. However, in our study, as *jcn* is not a normalised measure, *lin* is considered as the base line measure. Therefore, with comparing the *First/Second Order Similarity Measure* with *lin* measure, we can see that our proposed measure have higher Spearman's rank correlation.

6.2 Vector with bi-grams

The size of bi-grams all together is linked to four factors: 1) Corpus size, 2) Window size 3) Cut-off threshold for removal of low and high frequency bi-grams, and 4) Including or excluding of stop words. Usually large text plus lower window sizes result in

a better correlation. Pedersen et al. in their study [17] have shown that the number of bi-grams increases exponentially with the increase of the window size. It also represents the distribution of the Spearman's rank correlation for with different size of bi-grams. Generally, their representation shows that larger amount of bi-grams causes lower Spearman's rank correlation. In our study we changed the size of bi-grams (window size 2) in two ways, Stop words removal, and low and high frequency cut-off. Obviously, in both cases we would have the bi-gram size decreased.

For removing stop words, we used a list of 393 stop words and excluded them from bi-grams whenever they were encountered (whether as the first term in the bi-gram or second term). As the result of the elimination of stop words, these words would be removed from the constructed definitions of concepts used in the *Second Order Similarity Context Vector Measure* as well.

Table 3
Similarity Measures Results Before/After Stop Words Removal

Context Vector Similarity Methods	Before Stop Words Removal	After Stop Words Removal
1st Order Similarity	0.3357	0.3416
2nd Order Similarity	0.3265	0.3299

With applying stop words removal we not only could enhance the performance of our proposed methods in terms of speed, but also could achieve better results produced by both *First* and *Second Order Similarity Context Vector Measure*. These improvements in both cases are shown in table 3.

Pedersen et al. in their study [17] showed that using low and high frequency cut-off would increase the amount of correlation for measuring semantic relatedness between concepts in their method known as *vector* measure. With applying the idea of frequency cut-off (low frequency cut-off) in our study we reached to the results presented in table 4.

Table 4
Similarity Measures Results and Low Frequency Cut-off

	First Order Similarity	Second Order Similarity
No frequency cut-off	0.3416	0.3299
freq cut-off < 2	0.3374	0.325
freq cut-off < 5	0.3323	0.3203
freq cut-off < 10	0.3281	0.3163

The table 4 demonstrates the results of Spearman's rank correlation of *First* and *Second Order Context Vector* with comparing the original result (without considering any cut-off) and three low frequency cut-offs which are 2 or less, 5 or less and 10 or less. The findings indicate that applying low frequency cut-off does not improve the results in both *First* and *Second Order Similarity Measures* and in fact has adverse consequence.

6.3 Definition for Second Order Similarity Measure

In order to construct an extended definition for a specific biomedical concept in the *Second Order Context Vector Similarity Measure* numerous construction approaches are tested. One of these ways of definition construction in general outperforms others; therefore, that approach can be recommended for practical usages.

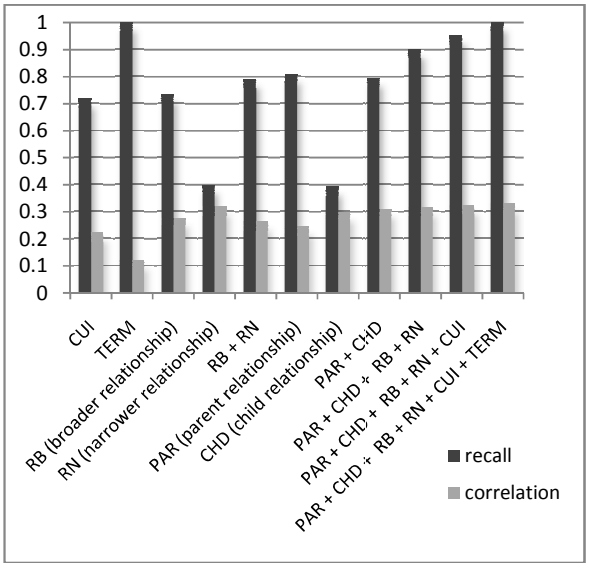

Fig.3 Concept Definition in Second Order Context Vector

This experiment acts on different approaches in order to augment definition of a concept (definition extension). These approaches are directly related to the various types of relationships in UMLS established among concepts.

Figure 3 represents the result of the experiment. It is shown that the best possible Spearman's correlation is achieved when extended definition is built based on full relationships, in other words "$PAR + CHD + RB + RN + CUI + TERM$". In this case the definition of each concept is augmented with the definition of its direct parents and children as well as the definition of its very next boarder and narrower concepts. The recall represents the percentages of how many pairs out of 301 pairs of concepts have definitions with different UMLS relations.

7 Discussion

The focus of the vector-based methods is the concept construction whether with considering all words in the corpus surrounding the concept or constructing an extended definition replaceable for concept. Both cases are based on the assumptions that the constructed definition from a thesaurus, as well as surrounding words of the concept from a corpus carry the same meaning that concept do. In other words, a high dimensional vector of an individual concept represents its distributional semantic. This format would help to extract tacit knowledge hidden in a larger corpus and exploit it to some appropriate extent.

In experience done in the study it is shown that *First Order Similarity Measure* outperform other method already proposed measures and our new method *Second Order Similarity Measure*. These methods exactness can be compared knowing correlation of their output and human judgments of concept pairs known as reference standard. Table 5 gives the three top and bottom pairs of concepts ranked by the First Order Context Vector Similarity.

Table 5
Semantic Similarity of 3 Top and Bottom Pairs of Concepts

Top 3 pairs	**Similarity**
emaciation / cachexia	0.960709
thalassaemias / Hemoglobinopathy	0.944226
Dyslipidaemia / hyperlipidaemias	0.938217
Bottom 3 pairs	**Similarity**
Metatarsalgias / Colchicines	0.002940
Meningisms / Acyclovir	0.002905
Hyperacuses / Bleomycins	0.002715

When comparing the information content vector based similarity results with the path similarity measure on the same dataset, the "path" measure [5] yields a much lower correlation (r=0.2913) than the ic-vector-based method (r=0.3416). Other path and depth based methods, proposed by Wu & Palmer [7] (r=0.2993) and Leacock & Chodorow [8] (r=0.2913) also have lower correlations than the *First* and *Second Order Similarity* methods. This is because path/depth-based approach relies exclusively on hierarchical relations.

Generally, findings of this research are useful for determining the appropriate semantics in the domain of biomedical. Since the measure of semantic similarity concept plays a crucial role in knowledge matching systems, the results provided by this research can be used to develop such matching, mapping and mediation systems applying semantic similarity algorithms as a fundamental part to resolve the data and knowledge heterogeneity problem. Furthermore, the proposed methods can be applied in different tasks such as information retrieval and word sense disambiguation. While these methods are independent from domain of study, their performance can be assessed in other specific domains as well.

8 Software Resources

The software for the similarity measures is part of UMLS::Similarity which is an open source software package [20] and can be downloaded from CPAN[5]. 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 [21]. The WordNet::Similarity package works as a foundation for 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[6] is to demonstrate the functionality of UMLS::Similarity without need the user to install the UMLS in MySQL database. It provides a way to introduce the package's source and relation.

In our study Purl codes of UMLS::Similarity and UMLS::Interface (an inter-connector between UMLS database and UMLS::Similarity package) are modified and included with modules of our methods to suit needs of the research.

9 Conclusions

This paper through introducing new methods for calculating semantic similarity draws a comparison between these new measures and other functions already proposed. The results of the study indicate that performance quality of a *Semantic Similarity Measure* very much relies on the function itself and the way it draws on the underlying information of the available resources. Similarly, the completeness of the resources and the fact that how much they can cover materials from the domain under study is vital

[5] CPAN: www.cpan.org

[6] http://atlas.ahc.umn.edu/cgibin/umls_similarity.cgi

for establishing the final similarity score. It denotes while proposed methods are fed with more accurate and richer information derived from appropriate resources the probability for having more reliable result increases. Following these rules, it is shown our proposed measures *First Order Context Vector Similarity Measure* and *Second Order Context Vector Similarity Measure* are more effective methods than other methods for semantic similarity.

References:

[1] S. Muthaiyah and L. Kerschberg, "A Hybrid Ontology Mediation Approach for the Semantic Web," *International Journal of E-Business Research*, vol. 4, 2008, pp. 79-91.

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

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

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

[5] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and application of a metric on semantic nets" *IEEE Transactions on Systems, Man and Cybernetics* vol. 19, 1989, pp. 17-30.

[6] J. Caviedes, and J. Cimino, "Towards the development of a conceptual distance metric for the UMLS." *Journal of Biomedical Informatics*, vol. 372, 2004, pp. 77-85.

[7] Z. Wu and M. Palmer, "Verb semantics and lexical selections" In *proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* 1994.

[8] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification in WordNet: An Electronic Lexical Database," 1998, pp. 265-283.

[9] J. Zhong, H. Zhu, J. Li and Y. Yu, "Conceptual graph matching for semantic search," *Proceedings of the 10th International Conference on Conceptual Structures*, pp. 92.

[10] H. A. Nguyen and H. Al-Mubaid, "New ontology-based semantic similarity measure for the biomedical domain," *In IEEE Eng Med Bio l Proc.,* 2006, pp. 623–628.

[11] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448-453.

[12] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," In *International Conference on Research in Computational Linguistics,* 1997.

[13] D. Lin, "An information-theoretic definition of similarity" In *15th International Conference on Machine Learning*. Madison, USA, 1998.

[14] S. Patwardhan and T. Pedersen, "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.

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

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

[17] T. Pedersen, Y. Liu, B. McInnes, G. Melton-Meaux and S. Pakhomov, "Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet," *Appears in the Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 2012, pp. 879.

[18] S. Banerjee and T. Pedersen, "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, 2002.

[19] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," *SIGDOC '86:Proceedings of the 5th annual international conference on Systems documentation*, New York, USA, 1986, pp. 24-26.

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

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