# Using SemRep to Label Semantic Relations Extracted from Clinical Text

Ying Liu[1], PhD, Robert Bill[1], Marcelo Fiszman[2], MD, PhD, Thomas Rindflesch[2], PhD,
Ted Pedersen[3], PhD, Genevieve B. Melton[1], MA, MD, Serguei V. Pakhomov[1], PhD
[1]University of Minnesota, Minneapolis, MN;
[2]National Library of Medicine, Bethesda, MD;
[3]University of Minnesota, Duluth, MN

## Abstract

*In this paper we examined the relationship between semantic relatedness among medical concepts found in clinical reports and biomedical literature. Our objective is to determine whether relations between medical concepts identified from Medline abstracts may be used to inform us as to the nature of the association between medical concepts that appear to be closely related based on their distribution in clinical reports. We used a corpus of 800k inpatient clinical notes as a source of data for determining the strength of association between medical concepts and SemRep database as a source of labeled relations extracted from Medline abstracts. The same pair of medical concepts may be found with more than one predicate type in the SemRep database but often with different frequencies. Our analysis shows that predicate type frequency information obtained from the SemRep database appears to be helpful for labeling semantic relations obtained with measures of semantic relatedness and similarity.*

## Introduction

Medical concepts are related to each other in a number of complex ways. The traditional way to identify these relations in biomedical literature is by using rule-based approaches[1]. For example, a tool designed for automatic identification of semantic predication from biomedical literature called SemRep[2, 3] operates by applying a set of linguistic rules to sentences found in Medline abstracts. Semantic relations identified by SemRep have been used in literature-based discovery (LBD) [4], among many other approaches to mining information from biomedical literature[5]. Biomedical articles on which SemRep is designed to operate contain explicit and implicit mentions of relationships between various medical concepts. For example a TREATS relation between a medication and a disorder may be found in a single sentence in a Medline citation containing the following text: “Metamorphosia associated with topiramate for migraine prevention.”

Clinical documents such as inpatient and outpatient clinical notes do not typically contain language that encodes the nature of the semantic relation. On the contrary, the TREATS association between “topiramate” and “migraine” is more likely to be found through co-occurrence between these two concepts in the same note as the two concepts are not likely to occur together in the same sentence or even section of the note. Clinical reports, however, constitute a rich source of empirical information on patient conditions and their treatment including information on potential medication allergies, side effects and adverse events. As such, clinical reports are an important source of complementary information and may be used to extract significant associations between biomedical concepts as they occur in practice on potentially large patient populations. Thus, we may use clinical notes to find concepts that are strongly related to each other, but we may not be able to determine the exact nature of the association from clinical documents. Conversely, biomedical literature provides a source of relationships between medical concepts that have been “distilled” through research. However, by nature of the publication process, information on associations between concepts available in biomedical literature is bound to be limited due to the inherent delay in conducting and publishing research. One potential way to leverage the strengths of both clinical text and biomedical literature is by mining the strength of associations between medical concepts from clinical reports and using biomedical literature to determine if the association found in clinical data has been studied and, if so, what is the most likely type of relationship for the association. This approach will potentially help in targeting interesting and clinically important associations (for example between medications and disorders) that have not been extensively examined before and may prove to signal either adverse drug reactions or lead to novel off-label uses of existing drugs.

Exploration of semantic relatedness between biomedical terms has become a popular topic in recent years. In our previous work, we developed a programmatic platform called UMLS-Similarity to automatically calculate the semantic similarity and relatedness for any pairs of concepts in the Unified Medical Language System (UMLS). Applications include Mathur et al. who used semantic similarity and relatedness public data set and UMLS-Similarity to calculate the gene and disease similarity[6]. Sahay et al. used the similarity and relatedness methods offered by UMLS-Similarity to connect relevant users together in conversation and to provide contextual recommendations relevant to the health information conversation system Cobot[7]. Ogren used the word-level semantic similarity and relatedness offered by UMLS-Similarity to improve the performance of the classifier OWCP (one-word conjunct pairs) where UMLS-similarity gave a 1.71% absolute increase in recall[8]. Semantic relatedness measures can also be used for improving automated acronym sense disambiguation[9] and measures of redundancy in clinical texts[10].

In this paper, we used UMLS-Similarity analysis to ascertain the relationship between semantic relatedness and the predicates of the SemRep database. We focused on the concept unique identifiers (CUIs) of the drugs and findings semantic groups, and used the relatedness scores to help us find the highly related CUI pairs which are not in the SemRep database.

## Background

### *SemRep*

The SemRep[2] database was developed at National Library of Medcine. It uses domain knowledge provided by the UMLS to represent the textual content as semantic predication. SemRep uses MetaMap[12] to map noun phrases to UMLS concepts. Through its rule-based summarization system, it maps the syntactic elements to semantic network predicates[13, 14]. About 117 millions of sentences are extracted from titles and abstracts of PubMed for the predication analysis. SemRep detects about 57 millions of predicate instances and 90 unique predicate types (about half of them are “NEG_predicate” such as “NEG_TREATS”).

### *Semantic Relatedness*

Methods for computing semantic similarity and relatedness are a class of computational technique. We follow Hirst and St-Onge[15] by treating semantic relatedness as a distinct and more general notion than semantic similarity. For the ontology-dependent methods, path-based methods are based on the path length between a pair of concepts in ontology[16, 17, and 18]. 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. Ontology-independent methods rely on distributional properties of concepts in large text corpora and may be easier to keep current with the changes in a given knowledge domain[19, 20].

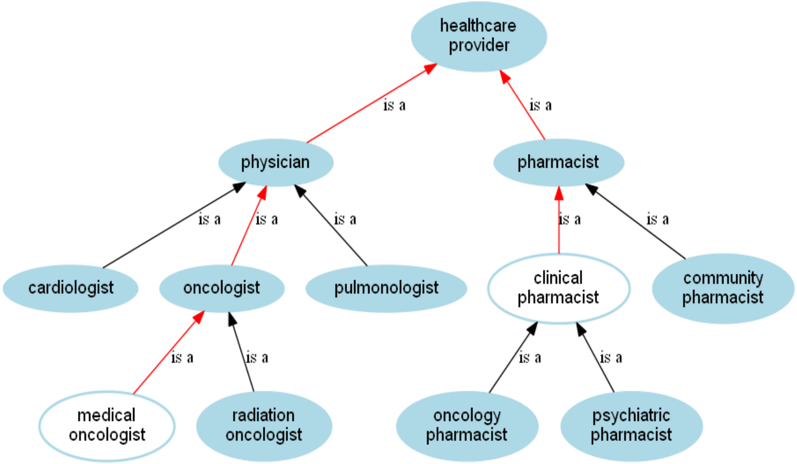

**Figure 1.** An example of the healthcare provider ontology.

In Figure 1, we present a simple healthcare provider ontology. According to the path-based measure, the similarity between medical oncologist and clinical pharmacist is 1/6. The similarity between medical oncologist and oncology pharmacist is 1/7. Although a medical oncologist and an oncology pharmacist are related due to their expertise in oncology, this is not reflected by the path-based method.

In this paper, we used the second-order context vector method to measure the semantic relatedness[21]. This is implemented and available in UMLS::Similarity[22]. For a pair of concepts with definitions, the second-order context vector method finds the context distribution of every word in the definition. The context distribution is recorded on a co-occurrence matrix constructed by scanning a large corpus. In Figure 2, the first order vectors recorded the semantic distribution. The second order vectors combined each word's first order vector together. The relatedness scores is the cosine of the two vectors. The relatedness score is from 0 to 1. 1 means two vectors are identical and 0 means two vectors are perpendicular to each other.

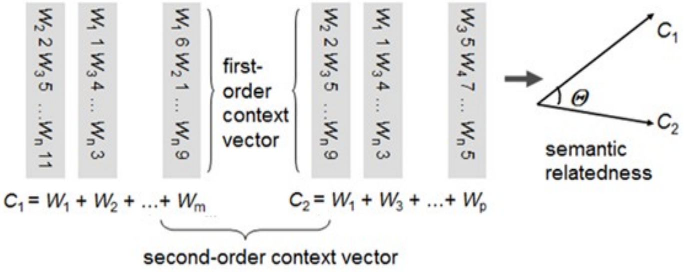
$W_2$ 2 $W_3$ 5 ...$W_n$ 11
$W_1$ 1 $W_3$ 4 ... $W_n$ 3
$W_1$ 6 $W_2$ 1 ... $W_n$ 9
first-order context vector
$W_2$ 2 $W_3$ 5 ...$W_n$ 9
$W_1$ 1 $W_3$ 4 ... $W_n$ 3
$W_3$ 5 $W_4$ 7 ... $W_n$ 5
$C_1$
$\Theta$
$C_2$
semantic relatedness
$C_1 = W_1 + W_2 + \ldots + W_m$
$C_2 = W_1 + W_3 + \ldots + W_p$
second-order context vector

**Figure 2.** Second order context vector method for semantic relatedness.

We used expanded relations in the UMLS in addition to WordNet to build comprehensive definitions. The results indicated that the method for extending concept definitions has the greatest effect of the performance of the context vector based approach. Table 1 lists some examples and their relatedness scores. From these examples we can see that high relatedness scores indicate either highly-related concepts or similar concepts. For medium related concepts, 'insulin' can treat 'diabetes and 'nocturia' is a symptom of 'diabetes'. Pairs with low relatedness scores mean they may not have strong relations. The data set used for this test is published by Pakhomov et al. and they are publicly available[23].

| Relatedness | Pairs |
|---|---|
| 0.92 | Warfarin<>Vitamin K, NOS |
| 0.87 | Dizzyness<>Vertigo NOS |
| 0.75 | Diabetes<>Insulin |
| 0.53 | Nocturia<>Diabetes |
| 0.31 | Pneumonia<>Weakness |
| 0.28 | Over nutrition<>Seizures, NOS |

**Table 1.** Semantic relatedness examples.

### Software Platforms

Our method relies on two software platforms. The first is UMLS::Similarity[1], which is an open source software package written in Perl that computes semantic similarity and relatedness between concepts in the UMLS. This is built on top of UMLS::Interface which interacts directly with the UMLS. UMLS::Similarity can be extended to include new measures and also includes a web interface[1]. The second software platform is BiomedICUS[2]. This is a Java open source software package which uses the Apache UIMA platform (Unstructured Information Management Architecture). BiomedICUS relies on another open source package known as uimaFit[3] which makes it easier to work with UIMA. BiomedICUS has an interface to MetaMap which helps to map the terms found in clinical reports to CUIs found in the UMLS.

[1] atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi
[2] http://code.google.com/p/biomedicus/
[3] http://code.google.com/p/uimafig/

## Experiments

### *Data*

We used a clinical notes corpus of 824,380 clinical notes from University of Minnesota-affiliated Fairview Health Services to compute semantic relatedness values. For labeling semantic relations extracted from Medline abstracts we use the SemRep database which contains 57 million predicates comprising 90 predicate types. Table 2 shows the top 12 predicates and their number of instances in the database. Among the 90 predicate types, we are particularly interested in the high frequency 'TREATS' and 'CAUSES' relations. These two relations are potentially confusable with each other. A drug can treat a symptom and a drug can also cause a symptom. Furthermore, the same pair of drug CUI and symptom CUI may be found by SemRep to be in a 'TREATS' and 'CAUSES' predicative relations.

| Predicate | # Instances |
|---|---|
| PROCESS_OF | 12,402,199 |
| LOCATION_OF | 9,278,433 |
| PART_OF | 8,533,663 |
| **TREATS** | **5,216,851** |
| ISA | 3,592,413 |
| COEXISTS_WITH | 2,416,926 |
| AFFECTS | 2,043,945 |
| INTERATS_WITH | 1,778,525 |
| USES | 1,306,120 |
| ASSOCIATED_WITH | 1,266,234 |
| **CAUSES** | **1,125,311** |
| ADMINISTERED_TO | 1,039,400 |

**Table 2.** Top 12 predicates and their instance frequencies in the SemRep databases (by May, 2012).

| | | | |
|---|---|---|---|
| **TREATS** | 1582 | AFFECTS | 20 |
| **CAUSES** | 10 | PREVENTS | 3 |
| NEG_TREATS | 2 | DISRUPTS | 1 |
| AUGMENTS | 1 | PREDISPOSES | 1 |

**Table 3.** Predicate types and number of instances for the subject C0040615 and object C0036341.

In Table 3, we provide an example showing the predicate types and their distribution for the pair of terms "antipsychotic agent" (C0040615) and "Schizophrenia NOS" (C0036341). These terms/concepts have eight types of predicates and a total of 1620 instances. The most frequent relation is 'TREATS' with 97.6% of the 1620 instances belonging to this predicate type, while the 'CAUSES' relation only covers 10 instances. In this particular case, if we were to find these two concepts to be closely semantically related based on in clinical reports, we could hypothesize that the most frequent TREATS relation is the most likely label for empirically determined association.

### *The Coverage of SemRep Predicates*

The experiment in this section addresses the coverage of the SemRep database of strong associations extracted from clinical text. We started with the clinical reports in XML format. The inpatient clinical notes were collected from 2003 to 2008 at Fairview Health Services. These semi-structured notes consist of admission history, physical operation, discharge summaries, and consultation notes. Thus the total size of the clinical notes corpus used in this study was about 209 million words. Our BiomedICUS system uses MetaMap (2010) to map each term to concept unique identifier (CUI). We selected those CUIs with higher than 900 mapping score and extracted the CUIs with the DRUG and FINDING semantic group (as shown in Tables 4 and 5).

The SemRep database records the predicates within one sentence. For those entities outside the same sentences, their relations are not recorded. In order to find the coverage of the 'TREATS' predicate, we constructed two semantic super-types DRUG and FINDING by grouping related subtypes (see Tables 4 and 5).

| | |
|---|---|
| **antb** | antibiotic |
| **horm** | hormone |
| **phsu** | pharmacologic substance |
| **orch** | organic chemical |
| **strd** | steroid |
| **vita** | vitamin |

**Table 4.** The DRUG semantic group.

| | |
|---|---|
| **dsyn** | disease or syndrome |
| **mobd** | mental or behavioral dsyfunction |
| **neop** | neoplastic process |
| **inpo** | injury or poisoning |
| **patf** | pathologic function |
| **anab** | anatomical abnormality |
| **sosy** | signs or symptom |
| **acab** | acquired abnormality |
| **cgab** | congenital abnormality |
| **comd** | cell or molecular |

**Table 5.** The FINDING semantic group.

We mapped the words in our clinical data repository of 824,380 notes to CUIs. For those CUIs with mapping scores higher than 900 and semantic types in the DRUG or FINDING groups, in total there were 7,374 CUIs in DRUG and 14,542 CUIs in FINDING. We used two methods to select CUIs from these two semantic groups. One is to sort the CUIs by their frequencies and then select the 1,000 CUIs with highest frequencies. The other method was to randomly sample 1,000 CUIs from the two semantic groups.

After we selected the CUIs, their semantic relatedness scores were calculated, forming two 1,000 by 1,000 CUI matrices (one for the random selection and one for the frequency-based selection), generating 1 million pairs of semantic relatedness calculations in each matrix. Subsequently, we checked each pair in each matrix to see if it existed in the SemRep database. The randomly selected 1,700 pairs are in the SemRep database; however, the high frequency set contained 33,656 pairs that were found in the SemRep database and 966,344 pairs that were not found. We focused on those pairs with high and medium relatedness scores that were not found in the SemRep database and conducted an informal analysis to identify potential reasons. For CUI pairs that were not found in SemRep database, we randomly selected pairs to perform a subsequent PubMed search using the preferred term for each CUI to determine if the concept pair can be found anywhere in Medline.

The results of this analysis are shown in Tables 6, 7, 8, and 9. We categorized the outcomes into the following broad categories:

(a) CUI pairs that occurred in SemRep (Table 6);

(b) CUI pairs that did not occur in SemRep but occurred in same sentence in at least one Medline abstract (Table 7);

(c) CUI pairs that did not occur in SemRep but occurred in the same Medline abstract but not in the same sentence (Table 8);

(d) CUI pairs that did not occur together either in SemRep database or the Medline abstracts (Table 9).

| Subject | Object | Predicates | Example Sentences | Relatedness |
|---|---|---|---|---|
| Simvastatin | Myeloma-Multiple | 3 TREATS | First clinical experience with simvastatin to overcome drug resistance in refractory multiple myeloma. (PMID=17655704) | 0.81 |
| Simvastatin | Lymphoma NOS | 1 TREATS<br>1 AFFECTS | Studies in severe combined immunodeficiency mice show that simvastatin delays the development of EBV-lymphomas in these animals. (PMID=15856040) | 0.76 |
| Simvastatin | Abnormal degeneration | 2 PREVENTS | We investigated whether simvastatin, a Food and Drug Administration-approved cholesterol-lowering drug, could protect against nigrostriatal degeneration after 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) intoxication to model PD in mice. (PMID=19864567) | 0.75 |

**Table 6.** Subject and object CUI occurred in the SemRep database.

| Subject | Object | Example Sentences | Relatedness |
|---|---|---|---|
| Metolazone | Vein Thrombosis, Deep | The usage of powerful diuretics, such as metolazone, may lead to thrombotic complications. (PMID= 3388002) | 0.80 |
| Muscle relaxant | Rigidity, muscle | Three cases of masseter muscle rigidity in the presence of nondepolarizing muscle relaxants were discovered. (PMID=10437710) | 0.85 |

**Table 7.** Subject and object CUI pairs that did not occur in SemRep
but occurred in same sentence in at least one Medline abstract

| Subject | Object | Example Sentences | Relatedness |
|---|---|---|---|
| Vincristine | Urinary Bladder Malignant Neoplasm | Urinary bladder cancer is one of the most common cancers worldwide. Human transitional cell carcinoma (TCC) cells are epithelial-like adherent cells originally established from a primary bladder carcinoma. Studies have shown that TCC cells are resistant to some chemotherapeutic agents such as vincristine (VCR) (PMID=2068935). | 0.88 |
| Hydrallazin | Aneurysm, NOS | Abdominal aortic aneurysm was induced by perfusion of an isolated aortic segment with elastase. Treatment with telmisattan (0.5 mg/kg per day) or hydralazine (15 mg/kg per day) was started after surgery and continued for 14 days. (PMID=19008714) | 0.79 |

**Table 8.** Subject and object CUI pairs that did not occur in SemRep
but occurred in the same Medline abstract but not in the same sentence.

| Subject | Object | Relatedness |
|---|---|---|
| Torsemide | Subdural hematoma, NOS | 0.80 |
| Torsemide | Colitis ischaemic | 0.81 |
| Metolazone | Varicosities | 0.84 |

**Table 9.** Subject and object CUI pairs that did not occur together either in SemRep database or the Medline abstracts.

## Discussion

The goal of this preliminary work was to investigate if explicit semantic relations between medical concepts extracted by SemRep from biomedical literature may be used to inform us as to the nature of strong semantic associations between medical concepts found through semantic relatedness analysis of clinical reports. In our previous work[11], we demonstrated that frequency of co-occurrence of drugs and disorders in a large corpus of clinical notes may be used to improve the precision of SemRep's extraction of TREATS relations. In this work, we examined the opposite direction and investigated the possibility of using SemRep to inform the process of automatic labeling of strong semantic associations extracted from clinical text.

An important result of this work is that we were able to find a substantial number of semantically closely related drug-finding concept pairs, as determined by their distribution in clinical reports, in the SemRep database. This means that SemRep can be used to find semantic labels motivated by biomedical research for these empirically determined associations. This is a promising result because it could potentially allow us to test whether a given closely semantically related drug-finding pair is already known to be in a TREATS or CAUSES relationship, thus informing our subsequent steps aimed at detecting and investigating drug safety signals. Not surprisingly, we also found a large number of strongly associated concept pairs in clinical reports that were not in the SemRep database.

Going forward, it will be important to investigate the reasons for this finding. The discrepancy between the contents of SemRep database and empirically mined associations may be indicative of gaps in medical knowledge and suggest new research targets for clinical and translational investigators to pursue. It may also help in the refinement and further development of SemRep rules and relations identification mechanisms. For example, our informal analysis of some of the concept pairs that were found to be associated in clinical reports but not in SemRep database suggests that exploring identification of semantic relations across a broader context spanning more than one sentence may be an important next step. Also, it may be important to examine relations contained in the full text of biomedical articles. Currently, the SemRep database used in our study only contained relations extracted from Medline citations. This discrepancy could also mean several other things including, poor concept matching, SemRep misses, as well as potentially purely spurious associations mined from clinical reports. These potential directions need to be investigated further. For example, one of the limitations of the approach used in the coverage study in this work is that we did not consider hierarchical relations between concepts when we tried to match drugs and findings from clinical reports to the SemRep database which does not contain brand name drugs, for example. Going forward it will be important to leverage hierarchical and other ontologic relations in order to improve the matching process. One of the possible solutions for this may be to use the UMLS-Similarity package to find synonymous and nearly synonymous concepts. However, from the standpoint of drug safety surveillance, the fact that we did not find a closely associated drug-finding pair in SemRep database may also be a useful signal in and of itself. It may mean that the empirically determined association based on reports generated in clinical practice may not have been scientifically investigated and may constitute an important but yet undiscovered adverse drug reaction. Furthermore, this information may be useful in Pharmacosurveillance 2.0 efforts aimed at finding new beneficial off-label uses for already approved medications.

### Conclusion

In this paper, we presented preliminary results of an exploratory study aimed at investigating the relationship between unlabeled semantic relations extracted from clinical reports and labeled relations extracted from biomedical literature. Our current findings indicate a strong potential for a synergistic relationship between the statistically driven methods for extracting strongly related concepts from clinical data and rule-based approaches such as SemRep for extracting predicates from biomedical literature. While very preliminary, our current finding hold promise for improving the use of large clinical text repositories for drug safety surveillance.

### Future Work

In future work, we plan to examine more in-depth the process of using SemRep for labeling strong associations mined from clinical reports. We would also like to expand the set of relations beyond the focus on drugs and findings. Other semantic types for which clinical reports will have rich distributional statistics information include allergies, symptoms, procedures and medical devices.

### Acknowledgements

This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine (Grant #R01 LM009623-01). We would like to thank Fairview Health Services for use of their clinical data.

### References

1. Nosofsk RM. Attention, similarity, and the identification–categorization relationship. 1986. Journal of Experimental Psychology: General, Vol 115(1), 39-57.
2. Rindflesch TC, Fiszman M. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Infor, 36: 462-77.
3. Ahlers CB, Fiszman M, Demner-Fushman D, Lang F, Rindflesh TC. 2007. Extracting semantic predication from MEDLINE citations for pharmacogenomics. In Pacific Symposium on Biocomputing, pp 209-220.
4. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. 2006. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc 2006, pp 349-353.
5. Wilkowski B, Fiszman M, Miller CM, Hristovski D, Arabandi S, Rosemblat G, Rindflesh TC. 2011. Graph-Based Methods for Discovery Browsing with Semantic Predications. AMIA Annu Symp Proc, pp 1514-1523.
6. Mathur S, Dinakarpandian D. 2012. Finding Disease Similarity Based on Implicit Semantic Similarity, Journal of Biomedical Informatics. 45(2), pp 363–371.
7. Sahay S, Ram A. 2011. Socio-Semantic Health Information Access. s. In Proceedings of the AAAI Spring Symposium on AI and Health Communication, Technical Report SS-11-01.
8. Ogren PV. 2011. Coordination Resolution in Biomedical Texts, Ph. D dissertation.
9. McInnes BT, Pedersen T, Liu Y, Melton GB, Pakhomov SV. 2011. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. 2011. AMIA Annu Symp Proc. pp 895-904.
10. Zhang R, Pakhomov SV, McInnes BT, Melton GB. Evaluating Measures of Redundancy in Clinical Texts. 2011. AMIA Annu Symp Proc. pp 1612–1620.
11. Rindflesch TC, Pakhomov SV, Fiszman M, Kilicoglu H, Sanchez VR. Medical facts to support inferencing in natural language processing. 2005. AMIA Annu Symp Proc. pp 634-8..
12. Aronson AR. 2001 Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annu Symp Proc, pp 17-21.
13. Fiszman M, Rindflesh TC, Kilicoglu H. 2004. Abstraction Summarization for Managing the Biomedical Research Literature. Proceeding CLS '04 Proceedings of the HLT-NAACL Workshop on Computational Lexical, pp 76-83.
14. Kilicoglu H, Fiszman M, Rodriguez A, Shin, D, Ripple AM, Rindflesch T. Semantic MEDLINE: A Web Application for Managing the Results of PubMed Searches. 2010. J Med Libr Assoc. 98(4): pp 273–281.
15. Hirst G, St-Onge D. 1998. Lexical chains as repre- sentations of context for the detection and correction of malapropisms. WordNet: An Electronic Lexical Database, pp 305–332.

16. Rada R, Mili H, Bicknell E, Blettner M. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems,Man, and Cybernetics, 19(1):17–30.
17. Wu Z, Palmer M. 1994. Verbs semantics and lexical selection. In Proceedings of the 32nd Meeting of Association of Computational Lin-guistics, pp 133–138.
18. Leacock C, Chodorow M. 1998. Combining local context and WordNet similarity for word sense identification. In WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, pp 265–83.
19. 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, pp 24–26.
20. Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. 2006. In Proceedings of the EACL workshop, Making sense of sense: bringing computational linguistics and psycholinguistics together. pp 1-8.
21. Liu Y, McInnes BT, Pedersen T, Melton-Meaux G, Pakhomov S. 2012. Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet. In the Proceedings of the 2nd ACM SIGHIT IHI , pp 363 – 371.
22. McInnes BT, Pedersen T, Pakhomov SV. 2009. UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. AMIA Annu Symp Proc, pp 431-435.
23. Pakhomov S, McInnes BT, Adam T, Liu Y, Pedersen T, Melton G. 2010. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. AMIA Annu Symp Proc, pp 572 – 576.