Research Report DTC 2005/12,
University of Minnesota Digital Technology Center, May 2005

Measures of Semantic Similarity and Relatedness in the Medical Domain

Ted Pedersen[1] **Serguei Pakhomov**[2] **Siddharth Patwardhan**[3]

[1] Department of Computer Science, University of Minnesota, Duluth, MN
[2] Division of Biomedical Informatics, Mayo Clinic, Rochester, MN
[3] School of Computing, University of Utah, Salt Lake City, UT

Abstract

In this paper we introduce a measure of semantic relatedness based on context vectors derived from medical corpora. We also extend a number of measures of semantic similarity for general English to the medical domain. We evaluate these methods with a newly created test bed of 30 medical concept pairs that were scored by three physicians and nine medical index experts. We find that our context vector measure correlates better with these human experts than do measures of similarity based on the medical ontology SNOMED-CT.

1 Introduction

A measure of semantic similarity takes as input two concepts, and returns a numeric score that quantifies how much they are alike, based on *is-a* relations. For example, *common cold* and *illness* are similar in that a *common cold* is a kind of *illness.* However, there are other relations between concepts such as *has-part, is-a-way-of-doing,* etc., that existing measures of similarity can not use since they only account for *is-a* relations.

This suggests that more general measures of semantic relatedness are needed to take advantage of increasingly rich ontologies (particularly in the medical domain) which have a wealth of relations beyond *is-a*. This is especially relevant in light of progress in automatically identifying a wide range of semantic relations in medical text (e.g., Rosario and Hearst, 2004). It seems likely that measures of relatedness such as we propose here can help to organize discovered relations between concepts, and thereby automatically augment existing ontologies with new relations and concepts.

Measures of semantic similarity and relatedness have also proven useful in a number of NLP tasks. For example, Budanitsky and Hirst (2001) identify malapropisms using various measures of similarity and relatedness. Resnik (1995), Patwardhan, et al. (2003), and McCarthy, et. al. (2004) employ measures of similarity in their approaches to word sense disambiguation. However, much this work has been relative to WordNet (Fellbaum, 1998), which focuses on general English concepts.

There are a growing number of ontologies that organize medical concepts into hierarchies and semantic networks, perhaps best exemplified by the Unified Medical Language System (UMLS) of the National Library of Medicine (NLM). The largest and most extensive of the ontologies included in UMLS is SNOMED-CT, which we use in the experiments in this paper. We adapted a number of measures of semantic similarity to SNOMED-CT, and compare those to our own corpus based measure.

In this paper we introduce a measure of semantic relatedness that derives context vectors from medical corpora. It is a robust measure since it does not rely on the structure of an ontology to measure relatedness between concepts. As such it can be used between any two concepts for which we have the necessary corpus based information (which will be described shortly). Here we confine its use to concept pairs that are joined in an *is-a* hierarchy (SNOMED-CT) since the measures to which we compare require this, but in general our measure is more flexible and does not require this.

This paper proceeds with a review of a number of existing measures of semantic similarity that have been applied to general English and the medical domain. Then we introduce the medical ontology SNOMED-CT and the Mayo Clinic corpus of clinical notes, which are our information sources

for the measures we explore in this paper. We introduce a new measure of relatedness based on second order context vectors derived from corpora. We also introduce a new test bed for the evaluation of measures of semantic similarity and relatedness in the medical domain. Finally, we present our experimental results, and suggestions for future work.

2 Measures of Semantic Similarity

Measures of semantic similarity are often based on information regarding *is-a* relations found in a concept hierarchy. This information can have to do with path lengths between concepts, or it may augment such structural information with corpus based statistics. We describe existing measures of similarity here, and then in Section 4 we introduce our own method, which is adapted from the word sense discrimination technique of Schütze (1998).

Path Finding Measures

When concepts are organized in a hierarchy, it is convenient to measure similarity according to structural measures that find path lengths between concepts. In fact, there have been a variety of such approaches proposed in both the medical domain and in general English.

Rada, et al. (1989) developed a measure based on path lengths between concepts in the Medical Subject Headings (MeSH) ontology[1]. They relied on *broader than* relations, which provide successively more or less specific concepts as you travel from concept to concept. They used this measure to improve information retrieval by ranking documents retrieved from MEDLINE, a corpus made up of abstracts of biomedical journal articles. More recently, Caviedes and Cimino (2004) developed a measure called CDist which finds the shortest path between two concepts in the UMLS. Their evaluation relative to a small set of concepts and concept clusters drawn from a subset of the UMLS consisting of MeSH, ICD9CM[2] and SNOMED shows that even such relatively simple approaches tend to yield reliable results.

Wu and Palmer (1994) present a measure of similarity for general English that relies on finding the most general concept that subsumes both of the concepts being measured. The path length from this shared concept to the root of the ontology is scaled by the sum of the distances of the concepts to the subsuming concept. Leacock and Chodorow (1998) define a similarity measure that is based on finding the shortest path between two concepts and scaling that value by twice the maximum depth of the hierarchy, and then taking the logarithm of the resulting score. Hirst and St-Onge (1998) introduce a path finding measure of relatedness, which is a more general relation than is similarity. In brief, the relatedness of two concepts is determined by the nature of the paths that join them; ideally this should be a path that is not too long and has relatively few changes in direction.

For the experiments in this paper, we developed a shortest path algorithm, and adapted the measure of Leacock and Chodorow for SNOMED-CT.

Information Content Measures

Resnik (1995) presents an alternative to path finding via the notion of *information content.* This is a measure of specificity assigned to each concept in a hierarchy based on evidence found in a corpus. A concept with high information content is very specific, while concepts with lower information content are associated with more general concepts. The information content of a concept is estimated by counting the frequency of that concept in a large corpus, along with the frequency of all the concepts that are subordinate to it in the hierarchy. The probability of a concept is determined via a maximum likelihood estimate, and the information content is the negative log of this probability.

Resnik defines a measure of similarity that holds that two concepts are semantically related proportional to the amount of information they share. The quantity of shared information is determined by the information content of the lowest concept in the hierarchy that subsumes both the given concepts. However, the Resnik measure may not be able to make fine grained distinctions since many concepts may share the same least common subsumer, and would therefore have identical values of similarity. Jiang and Conrath (1997) and Lin (1998) developed measures that scale the information content of the subsuming concept by the information content of the individual concepts. Lin does this via a ratio, and Jiang and Conrath with a difference.

[1] MeSH is distributed by the National Library of Medicine
[2] International Classification if Diseases, 9th revision, Clinical Modification

Lord, et al. (2003) adapted these three information content measures to the Gene Ontology (GO). They found that these measures can be successfully used for "semantic searching" of the textual resources available to bioinformatics research.

We adapted the measures of Resnik, Lin, and Jiang and Conrath to SNOMED-CT, using the Mayo Clinic clinical notes for information content estimates.

3 SNOMED-CT

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

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

One characteristic of SNOMED-CT that presents challenges for calculating semantic similarity is that it allows multiple inheritance where a concept can belong to more than one of the 13 hierarchies. As a preliminary solution, we introduce a "root" node that is the parent of the top 13 hierarchies and the path length between two concepts is calculated by taking the average path lengths for each hierarchy in which the concept was found.

4 Mayo Clinic Corpus of Clinical Notes

The corpus that was used in this paper consists of ~1,000,000 clinical notes which cover a variety of major medical specialties at the Mayo Clinic. Clinical notes have a number of specific characteristics that are not found in other types of discourse, such as news articles of even scientific medical articles found in MEDLINE. Clinical notes are generated in the process of treating a patient at a clinic and normally represent the dictations every physician practicing in the US is required to file. As a result, these notes represent a kind of quasi-spontaneous discourse (Anonymous) where the dictations are made partly from notes and partly from memory. More often than not, the speech tends to be telegraphic which presents certain challenges for natural language processing.

****CC****
Review recent progress.
****CM****
Aspirin 81 mg q.d.
Imdur 30 mg q.d.
Lisinopril 5 mg q.d. (increased to 10 mg q.d. today)
****HPI****
Her vocal cord examination yesterday was unremarkable. She broke her ankle toward the end of YEAR and is still limping but it is improving. While she was hospitalized for aspiration pneumonia after her vocal cord biopsy in DATE, she developed tachycardia with ECG changes. Echocardiogram showed EF of 30-35% with regional wall motion abnormalities. She was started on Lisinopril and Imdur.
****IP****
#1 Probable CAD
#2 ASO
Plan: Because of some elevated blood pressure, we will increase her Lisinopril to 10 mg q.d.
****SI****
DISM 1/13/99
****DX****
#1 Probable CAD
#2 ASO

Figure 1. A short excerpt from a Clinical Note

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

(DX), among others. A typical example of a clinical note is given in Figure 1.

We are particularly interested in the CC, HPI, IP and DX section of the clinical notes. The CC section records the reason for visit; HPI section has information of what other treatments/problems the patient has had in the past; IP section contains the diagnostic and current treatment information, while the DX section is an abstraction of the IP section – it contains only a list of diagnoses. Other sections such as SI (Special Instructions) and CM (Current Medications) are less interesting from the standpoint of semantic relatedness measures, although if we were to focus on computing semantic relatedness between medications, then we may want to consider the CM section as well.

5 Context Vector Measure

We have developed a measure of semantic relatedness that represents a concept with a context vector. This is more flexible than measures of similarity since it does not require that the concepts be connected via a path of relations in an ontology.

Our context vector measure is an adaptation of Schütze's (1998) method of word sense discrimination. First, a co-occurrence matrix is constructed such that each cell contains the log-likelihood score between a term found in the description of a concept, and each of the words it co-occurs with in a given corpus. Thus, the rows of this matrix represent the description terms that are used to define concepts, while the columns are the words with which it occurs in a large corpus of text. The exact nature of the description words for a concept can vary, but could consist of a gloss or definition, or a set of related words.

We have created context vectors that rely on a rich source of descriptions that has been systematically collected over the past ten years at Mayo and represents a large amount of human coding experience. This resource contains over 16 million unique diagnostic phrases expressed through natural language that correspond to over 21,000 diagnoses and represents an utterance level thesaurus. This database provides entry points to patient information at Mayo Clinic. Each diagnostic statement has been uttered by a practicing physician at the Mayo Clinic as part of the patient's medical record manually coded and cataloged for subsequent retrieval using a Mayo Clinic modified Hospital International Classification of Diseases Adaptation (HICDA). The HICDA classification is a hierarchy consisting of four levels. The top level is the most general and has 19 categories such as *Neoplasms, Diseases of the Circulatory System,* etc. The next 3 levels group diagnoses into more specific categories.

The Mayo Clinic thesaurus is constructed on the assumption that if several diagnostic phrases have been classified to the same category in the HICDA hierarchy, then these phrases can be considered as synonymous at the level of granularity afforded by HICDA. For example, diagnostic phrases such as "primary localized hilar cholangiocarcinoma" and "cholangiocarcinoma of the Klatskin variety" are linked in a thesaurus-like fashion because these two statements have been manually classified the same way. We consider these two phrases nearly synonymous and use them to generate quasi-definitions for terms found in both SNOMED and this utterance level thesaurus of diagnostic phrases. There is a fair amount of noise in this collection, which we attempt to reduce by excluding those phrases that occur 5 times or less and those phrases that are classified as "*Admission, diagnosis not given.*" The Mayo Clinic thesaurus is then merged with the UMLS. The merging consists of string-matching each term in the Mayo Clinic thesaurus to a concept in the UMLS and then transferring all the other terms associated with the UMLS (which subsumes SNOMED) concept into the Mayo Clinic thesaurus. The resulting thesaurus contains 3,665,721 diagnostic phrases organized into 594,699 clusters, which is an average of 6 terms in a cluster. The thesaurus represents 344,550 or 95% of SNOMED-CT concepts.

Then, to represent the concepts that occurred in both the thesaurus and SNOMED-CT for semantic similarity measures, we take each of the descriptor terms in the thesaurus/cluster and build a co-occurrence matrix for it from the Mayo Clinic clinical notes. After the vectors are created, the concepts represented by the descriptor words are themselves represented by an averaging of all the vectors associated with all the descriptor words. Thus, a SNOMED-CT concept is represented entirely outside of SNOMED-CT by way of an averaged vector of word co-occurrences, where the words that represent a concept are derived from an automatically created thesaurus.

6 Experimental Data

Measures of semantic similarity can be evaluated both directly and indirectly. The direct method compares systems relative to human judgments; common standards for general English are provided by Rubenstein and Goodenough (1965) and Miller and Charles (1991). The indirect methods evaluate similarity and relatedness measures based the performance of the application that relies on the measures. Spelling correction (Budanitsky and Hirst, 2001) and word sense disambiguation (Patwardhan, Banerjee, and Pedersen, 2003) have been shown to be sensitive to semantic relatedness.

For the medical domain there are no existing sets of related words that have been created by human experts that could be used in our study. As such we created a test bed of pairs of medical terms that were scored by human experts according to their relatedness. We asked a Mayo Clinic physician trained in Medical Informatics to generate a set of 120 term pairs representing a range of relatedness from *not related at all* to *very closely related*. The number 120 reflects the fact that we asked to generate pairs in four broad categories, following Rubenstein and Goodenough, with 30 term pairs in each. Subsequently, we had 13 medical index experts annotate each pair with a relatedness value on a scale of 1 to 10. The Medical Index consists of a group of people who are trained to classify clinical diagnoses using the same HICDA classification system used in the Mayo Clinic thesaurus.

The classification is done primarily for subsequent identification of patient cohorts that match the criteria requested by health science research investigators who conduct epidemiological studies at the Mayo Clinic. The experts who annotated the test set for this study have had between 5 and 14 years of coding experience. Although they do not have formal training in medicine, by virtue of working with clinical records and terminologies have had a lot of exposure to medical language and we considered them as good candidates for this annotation task.

	W1	W2	Phys.	Expert
4	renal failure	kidney failure	4.0000	4.0000
5	heart	myocardium	3.3333	3.0000
1	stroke	infarct	3.0000	2.7778
7	abortion	miscarriage	3.0000	3.3333
9	delusion	schizophrenia	3.0000	2.2222
11	congestive heart failure	pulmonary edema	3.0000	1.4444
8	metastasis	adenocarcinoma	2.6667	1.7778
17	calcification	stenosis	2.6667	2.0000
10	diarrhea	stomach cramps	2.3333	1.3333
19	mitral stenosis	atrial fibrillation	2.3333	1.3333
20	***chronic obstructive pulmonary disease***	***lung infiltrates***	***2.3333***	***1.8889***
2	rheumatoid arthritis	lupus	2.0000	1.1111
3	brain tumor	intracrainial hemorrhage	2.0000	1.3333
15	carpal tunnel syndrome	osteoarthritis	2.0000	1.1111
18	diabetes mellitus	hypertension	2.0000	1.0000
27	acne	syringe	2.0000	1.0000
12	antibiotic	allergy	1.6667	1.2222
13	cortisone	total knee replacement	1.6667	1.0000
14	pulmonary embolus	myocardial infarction	1.6667	1.2222
16	pulmonary fibrosis	lung cancer	1.6667	1.4444
6	cholangiocarcinoma	colonoscopy	1.3333	1.0000
29	lymphoid hyperplasia	laryngeal cancer	1.3333	1.0000
21	multiple sclerosis	psychosis	1.0000	1.0000
22	appendicitis	osteoporosis	1.0000	1.0000
23	rectal polyp	aorta	1.0000	1.0000
24	xerostomia	alcoholic cirrhosis	1.0000	1.0000
25	peptic ulcer disease	myopia	1.0000	1.0000
26	depression	cellulitis	1.0000	1.0000
28	varicose vein	entire knee meniscus	1.0000	1.0000
30	hyperlipidemia	metastasis	1.0000	1.0000

Figure 2. Test set of 30 medical term pairs sorted in the order of the averaged physicians' scores.

As a control, we had 10 of the 13 experts[3] annotate the 30 general English term pairs in the Rubenstein and Goodenough's and Miller and Charles' test set. This was done to make sure the experts understood the instructions and the notion of relatedness. The correlation of the medical index experts' judgments with those of the annotators used by Rubenstein and Goodenough was relatively high – 0.84. Similarly, the correlation with the Miller and Charles's test set was even better – 0.88. Unfortunately, the agreement on the medical test set of 120 concept pairs turned out to be fairly low - 0.51. In order to derive a more reliable test set we extracted only those pairs for which the agreement was relatively high. This resulted in a set of 30 concept pairs (displayed in Figure 2) which were then annotated by three physicians and a subset of 9 available medical index experts from the 13 who annotated the original 120 pairs. All three physicians are specialists in the area of rheumatology. The fact that all of them specialize in the same sub-field of medicine can be helpful in get-

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

ting good inter-rater agreement. Each pair was annotated on a 4 point scale: "*practically synonymous, related, marginally related and unrelated.*" We have listed the term pairs and the averaged scores assigned by the physicians and the experts in Figure 2. The term pair 20 in bold has been excluded from the test bed because the term "lung infiltrates" was not found in the SNOMED-CT terminology. Thus, the resulting test set consists of 29 pairs; however, we were able to calculate the inter-rater correlation using all 30 pairs. The average correlation between physicians is 0.68. The average correlation between experts is 0.78. We also computed the correlation across the two groups after we averaged the scores each member of the respective groups had assigned to each pair in the test set. The correlation across groups is 0.85.

7 Experimental Results

We implemented two measures of semantic similarity based on the structure of SNOMED-CT: the shortest path algorithm and the Leacock and Chodorow measure. We implemented three measures that are based on a combination of information content statistics derived from the Mayo Clinic clinical notes, and the ontology provided by SNOMED-CT. Finally, we implemented our own context vector measure by finding co-occurrence vectors from the Mayo Clinic clinical notes based on the descriptor terms associated with concepts in the Mayo Clinic thesaurus. As such the context vector measure was the only measure that was not based on SNOMED-CT in some way.

First, we tested two hypotheses related to the context vector measure. The first was that the correlation between the measure and the experts depends on the size of the training corpus. The second hypothesis was that the section type of the clinical notes also has an effect on the correlation. In the following sections, we test these two hypotheses and show a comparison between all available measures of semantic relatedness.

Corpus Size

We experimented with training the Context Vector measure on variable amounts of text ranging from a corpus of 100,000 clinical notes to 1M notes.

Number of notes	Matrix size	N tokens
100K	32594X32594	21,593,156
200K	43179X43179	43,459,602
300K	50928X50928	66,176,995
400K	57328X57328	88,885,380
500K	62733X62733	111,019,453
600K	64910X64910	133,719,224
700K	67382X67382	156,467,734
800K	69883X69883	179,114,059
900K	72911X72911	206,489,197
1000K	75195X75195	232,080,038

Table 3. Descriptive training corpus statistics. (frequency range = 5-1000)

In order to build the matrices, we set the parameters so that we would only count the co-occurrence of terms that occurred more than 5 times and less than 1000 times. Table 3 provides descriptive statistics on the corpus size and the resulting co-occurrence matrix size.

Test results relative to the test set of 29 term pairs are shown in Figure 3. The overall trend suggests that the correlation between relatedness judgments of the context vector measure and those of human experts improves with larger amounts of training data where 300K size appears to be the point where gains level off.

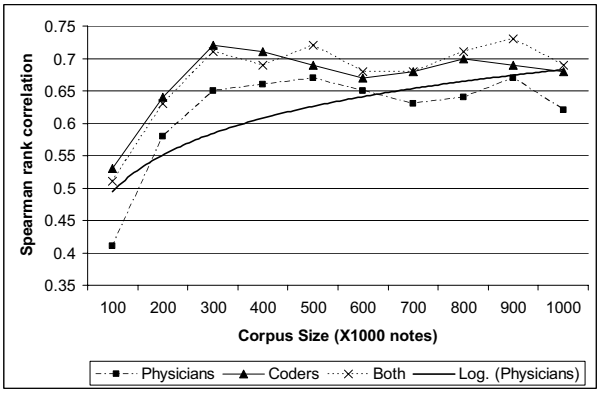

Figure 3. Correlation with human experts for the context vector measure trained on all sections. The log trend line is showing improvement with corpus size.

Figure 3 shows correlation with the scores provided by the group of physicians and the experts separately as well as combined scores averaged across both groups. For these experiments, we used data from 4 sections of the clinical notes - Chief Complaint (CC), History of Present Illness (HPI), Impression/Plan (IP) and Final Diagnosis (DX).

Section Type

We then experimented with the four section types by keeping the corpus size constant at 100K notes and varying the section type while building the context vectors. Table 4 displays the results.

section	physicians	experts	both	N tokens
IP	**0.56**	**0.59**	**0.6**	10,883,117
CC	0.47	0.53	0.53	956,438
DX	0.53	0.55	0.56	490,417
HPI	0.46	0.54	0.56	7,487,209
ALL	0.41	0.53	0.51	21,593,156

Table 4. Correlation results for the context vector measure trained on different sections of a 100K notes corpus.

The best correlation is achieved on the corpus compiled from the IP sections, closely followed by DX. This is not surprising as the IP section contains the diagnostic information pertinent to the patient's condition and intuitively should contain more closely related terms than other sections. The DX section is an abstraction of the IP section in that it only contains the diagnoses without additional descriptions.

We conclude that the most optimal context vector measure would result from training on the IP sections of the entire 1M notes corpus. In the following section, we compare the most optimal context vector measure to other measures.

Comparison with other measures

Table 5 shows the results of comparing all available measures of semantic relatedness.

Method	Phys.	Expert	Both
VECTOR (IP only, 1M notes)	0.84	0.75	0.76
VECTOR (All sect, 1M notes)	0.62	0.68	0.69
Lin	0.6	0.75	0.69
JCN	0.45	0.62	0.55
RES	0.45	0.62	0.55
PATH	0.36	0.51	0.48
LCH	0.35	0.5	0.47

Table 5. Comparison of correlations across measures with context vector measure trained on the IP sections only of 1M notes.

The vector based measure appears to produce better results than any other measure (by a wide margin where physicians are concerned), followed by the information content based measures and then the path based measures. This is an interesting result as it suggests that an ontology-independent measure can perform at least as well or better than the best of the ontology-based measures. Another interesting observation is that context vector measure produces a much closer correlation with physicians that with experts. For all other measures, this is reversed. We hypothesize that this result is due to the nature of the professional training and activities of the two groups – experts are trained in using hierarchical classifications, while physicians are trained to diagnose and treat patients. One possible indication from this observation is that the data contained in the clinical notes may reflect certain kinds of semantic relations between medical concepts in the mind of a physician better than a handcrafted medical ontology such as SNOMED-CT. By all means, more experimentation is necessary in order to test this hypothesis.

It is also worth pointing out that the context vector based measure trained on the IP sections of 1M notes performs considerably better than the measure trained on all sections of 1M notes, especially on physician's judgments as shown in the first two rows of Table 5.

8 Conclusions and Future Work

The existence of semantic equivalence classes between lexical items in English makes it highly desirable to use thesauri of synonymous or nearly synonymous terms for information (IR) and document retrieval (DR) applications. The issue is particularly acute in the medical domain due to stringent completeness requirements on such IR tasks as patient cohort identification. An epidemiologist performing an incidence study would rather sift through irrelevant patient records than miss any potentially relevant patients. We believe that semantic relatedness can improve the performance of such systems, since being able to map the user's search query for "congestive heart failure" to include *cardiac decompensation*, *pulmonary edema, ischemic cardiomyopathy* and *volume overload* as terms related to *congestive heart failure*. Clearly, *pulmonary edema* does not denote the same or even a similar disorder as *congestive heart failure* but under the patient cohort identification conditions it could be considered as an equivalent search term.

In our experiments we have been able to show the efficacy of adapting WordNet based semantic relatedness measures developed for general Eng-

lish to a specialized subdomain of biomedicine represented by SNOMED-CT. We have also determined that an ontology-independent context vector measure is at least as good as other ontology-dependent measures, provided that there is a large enough corpus of unlabeled training data available. This finding is important because developing specialized ontologies such as WordNet, SNOMED or UMLS is a very labor intensive process. Also, there are some indications that manually constructed ontology may not fully reflect the reality of semantic relationships in the mind of a practicing physician. The vector based measure can help alleviate these problems in addition to the benefit of rapid adaptation to a new domain.

In the near future, we plan to extend the measures of relatedness to use the UMLS as a source of the ontological relations for path-based measures as well as glosses for vector based measures. A fairly complete list of definitions is provided in the latest version of UMLS 2004AC. We also would like to experiment with applications of semantic relatedness measures to NLP tasks such as word-sense discrimination, information retrieval and spelling correction.

Acknowledgements

This work was supported in part by a grant from the Digital Technology Center (DTC) of the University of Minnesota under their Digital Technology Initiative (DTI) program in 2004-2005.

References

Banerjee, S. and Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence; pages 805-810; August; Acapulco, Mexico.

Budanitsky, A. and Hirst G. (2001). Semantic Distance in WordNet: An Experimental Application Oriented Evaluation of Five Measures; Proceedings of the Workshop on WordNet and other Lexical Resources: Applications, Extensions, and Customizations; pages 29-34; June; Pittsburgh, PA.

Caviedes, J. and Cimino, J. (2004) Towards the development of a conceptual distance metric for the UMLS. Journal of Biomedical Informatics 37: 77-85.

Chute CG, Crowson DL, Buntrock JD. (1995) Medical information retrieval and WWW browsers at Mayo. Proceeding of Annual Symposium on Comput. Appl. Med. Care, pages 903-907.

Fellbaum, C. (Ed.) (1998) WordNet: An Electronic Lexical Database. MIT Press. Cambridge, MA.

Hirst, G. and St-Onge, D. (1998) Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms; In: Fellbaum, C. (Ed.) (1998) WordNet: An electronic lexical database; pages 305-332. MIT Press. Cambridge, MA.

Jiang, J. and Conrath, D. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy; Proceedings of International Conference on Research in Computational Linguistics; pages 19-33; Taipei, Taiwan.

Leacock, C. and Chodorow, M. (1998) Combining Local Context and WordNet Similarity for Word Sense Identification; In: Fellbaum, C. (Ed.) (1998) WordNet: An electronic lexical database; pages 265-283; MIT Press; Cambridge, MA.

Lin, D. (1998) An Information Theoretic Definition of Similarity; In Proceedings of the 15th International Conference on Machine Learning; pages 296-304; Madison, WI.

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. Bioinformatics, **19(10):**1275-83.

McCarthy, D., Koeling, R., Weeds, J. and Carroll, J. (2004) Finding predominant senses in untagged text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain. pp 280-287

Miller, G. and Charles, W. (1991) Contextual Correlates of Semantic similarity; Language and Cognitive Processes; **6(1)**:1-28.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003) Using Measures of Semantic Relatedness for Word Sense Disambiguation; Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics; pages 241-57; February; Mexico City;

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.

Resnik, P. (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy; Proceedings

of the 14th International Joint Conference on Artificial Intelligence; pages 448-453; August; Montreal.

Rosario, B. and Hearst, M. (2004) Classifying Semantic Relations in Bioscience Texts. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics; pages 430-437; Barcelona, Spain.

Rubenstein, H. and Goodenough, J.B. (1965) Contextual Correlates of Synonymy; Computational Linguistics; **8:**627-633.

Schütze, H. (1998). Automatic Word Sense Discrimination; Computational Linguistics; **24 (1)**: 97-123.

SNOMED-CT: Fact Sheet. Available from http://www.snomed.org/snomedct/documents/July04_CTFactSheet.pdf

Wu, Z. and Palmer, M. (1994) Verb Semantics and Lexical Selection; Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics; pages 133-138; Las Cruces, NM.

UMLS: Unified Medical Language System. Available from http://www.nlm.nih.gov/research/umls/.