

Resolving Ambiguities in Biomedical Text With Unsupervised Clustering Approaches

Guergana Savova¹, PhD, Ted Pedersen², PhD,
Amruta Purandare³, MS, Anagha Kulkarni², BEng

¹Biomedical Informatics Research, Mayo Clinic, Rochester, MN

²Computer Science Department, University of Minnesota, Duluth, MN

³Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA

May 10, 2005

Abstract

This paper explores the effectiveness of unsupervised clustering techniques developed for general English in resolving semantic ambiguities in the biomedical domain. Methods that use first and second order representations of context are evaluated on the National Library of Medicine Word Sense Disambiguation Corpus. We show that the method of clustering second order contexts in similarity space is especially effective on such domain-specific corpora. The significance of the current research lies in the method extension to a new, previously untested domain and the general exploration of method portability across domains.

1 Introduction

One of the most important problems in biomedical text processing is associating terms that appear in corpora with concepts that are known in ontologies, such as the Unified Medical Language System (UMLS), developed at the National Library of Medicine (NLM) of the National Institute of Health (NIH)¹. Such mappings can help analyze medical text for semantic-based indexing and retrieval purposes, as well as build decision support systems for the biomedical domain. Word sense disambiguation is among the most significant challenges in mapping terms to a given ontology, which is necessary when a given term maps to more than one possible concept or sense.

The impact of semantic ambiguity in biomedical text processing is well documented. For example, Weeber et al., (2001) observed that the main source of errors in the NLM Indexing initiative was related to semantic ambiguity. This initiative seeks to “investigate NLP methods whereby automated indexing techniques can partially or completely substitute for current (manual) indexing practices” used for the retrieval of biomedical literature. A study of the UMLS Metathesaurus reported more than 7,400 ambiguous strings that map to more than one thesaurus concept (Roth & Hole, 2000). Similarly, Friedman (2000) described the challenges in this area in her group’s efforts to extend Medical Language Extraction and Encoding System (MedLEE), which is used for automated encoding of clinical information in text reports into Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) and UMLS codes. Chen, Liu and Friedman (2005) investigate the extent of gene name ambiguity in a set of biomedical publica-

¹ <http://www.nlm.nih.gov/pubs/factsheets/umls.html>

tions and report ambiguity rates as high as 85% which seriously affects the appropriate identification of gene entities.

Clearly there is a need for improved capabilities in resolving semantic ambiguity in biomedical texts. The dominant approach in word sense disambiguation is based on supervised learning from manually sense-tagged text. While this is effective, it is quite difficult to get a sufficient number of manually sense-tagged examples to train a system. Mihalcea (2003) estimates that 80-person years of annotation would be needed to create training corpora for 20,000 ambiguous English words, given 500 instances per word. Similar problems of scale exist for creating manually sense tagged text for the biomedical domain. In addition, the dynamic nature of biomedical text demands that we seek solutions that are not locked into a particular set of meanings, or require very extensive hand built knowledge sources to pursue.

For these reasons we are developing unsupervised knowledge-lean methods that avoid the bottlenecks created by sense-tagged text. Unsupervised clustering methods only utilize raw corpora as their source of information, and there are growing amounts of specialized biomedical corpora available.

This paper is organized as follows. First, we review previous work in this area that motivates our current approach, and then describe the approach in detail. Next, we describe our experimental data, which is a sense-tagged corpus made available from the NLM. Then, we present our experimental results. We close with an overview of future work.

2 Relation to Previous Work

A number of knowledge-lean unsupervised approaches have been developed for discovering and distinguishing among word senses in general English (e.g., Pedersen & Bruce, 1997; Schütze, 1998; Pantel & Lin, 2002; Purandare & Pedersen, 2004). In these studies the discovered clusters are evaluated by comparing them to sense distinctions made in a general English dictionary, or by evaluating how well the method distinguishes between unambiguous words that were conflated together as a pseudo-word.

However, the sense distinctions of interest in biomedical text are often relative to an ontology that not only acts as a dictionary, but has much broader applications. The reliance on ontologies in biomedical text processing is advantageous in that the structure of the ontology can be used to assign senses (e.g., Widdows et al., 2003). However, biomedical ontologies in general, and UMLS in particular, are constantly evolving and there are often senses that must be added. Thus, resolving ambiguity in the biomedical domain includes not only the traditional task of assigning previously determined senses to terms, but also recognizing new senses that are not yet a part of the ontology.

Liu, Lussier and Friedman (2001) also point out that disambiguation in the biomedical domain is distinct from general English since the nature of the sense distinctions and their granularity may be significantly different.

This paper seeks to extend existing methods of unsupervised word sense discrimination to biomedical text. It adopts the experimental framework proposed by Purandare and Pedersen (2004). They created three systems that follow Schütze (1998) and use second order co-occurrences as the main source of information. They also created three systems that rely on first order features, following Pedersen and Bruce (1997). The goal of Purandare and Pedersen was not to replicate these specific earlier methods, but rather to model broad classes of unsupervised methods and put them in a framework that allowed for convenient and systematic comparison.

Purandare and Pedersen (2004) compared the effectiveness of these six methods by discriminating among the meanings of target words drawn from the Senseval-2 corpora, and the *line*, *hard*, and *serve* corpora. As such, it is fair to say that these methods have only been evaluated on general English. We seek to investigate the portability and extension of those methods across domains, thus laying the background for method improvements and more general understanding of their strengths and weaknesses.

3 Unsupervised WSD Methodology

The goal of our method is to divide the contexts that contain a particular target word into clusters, where each cluster represents a different meaning of that target word. Each cluster is made up of similar contexts, and we presume that a target word used in similar contexts will have the same or very similar meaning. This allows us to discover word senses without regard to any existing knowledge source and remain completely unsupervised.

The data used in this study consists of a number of contexts that include a given target word, where each use of the target word has been manually sense tagged. The sense tags are not used during clustering; rather they provide a means of evaluating the discovered clusters. As such the methods here are completely unsupervised, and the sense tags are not used for either feature identification or for clustering.

The contexts to be clustered are assumed to be our only source of information about the target words, so the features used during clustering are identified from this very same data. Note that the sense tags mentioned above are not included in the data when features are identified.

Our goal is to convert the contexts into either a first or second order context vector. First order vectors directly represent the features that occur in a context to be clustered. Second order feature vectors are simply the average of several first order vectors. Both first and second order feature vectors can be clustered directly using their vector forms (vector space clustering) or by computing a similarity matrix that shows pair-wise similarities among the contexts (similarity space clustering).²

Clustering continues until a pre-specified number of clusters are found. This either corresponds to the number of senses present in the data, or a hypothesized number of clusters. We would prefer not to set the stopping point for the clustering, and are working on methods that will automatically stop at an appropriate point.

The discovered clusters are evaluated by comparing them to the manually assigned sense tags, which have been completely removed from the process until the evaluation.

Lexical Features

All of the methods in this study rely on lexical features to represent the context in which a target word occurs. These features include bigrams, co-occurrences, and target co-occurrences. Bigrams are ordered pairs of words that occur within five positions of each other five times or more. Note that this means there can be up to three intervening words between them. In addition to frequency, we require that the words in the bigram have a Log-likelihood ratio of more than 3.841. This indicates that there is a 95% chance ($p\text{-value} < 0.05$) that the two words are statistically dependent (Dunning 1993). Co-occurrences are identical to the bigram features, except they are unordered. Target co-occurrences are simply co-occurrences that include the target word.

First Order Methods

² All experiments reported here are performed using the SenseClusters package, <http://senseclusters.sourceforge.net>

The first order methods are loosely based on Pedersen and Bruce (1997) and for that reason are referred to as PB1, PB2, and PB3. The PB methods are all based on first order context vectors, which means that this vector directly indicates which features occur in that context. These methods use target co-occurrences or bigram features as described in the previous section. A similarity matrix or the actual context vectors are clustered using the average link agglomerative method or repeated bisections (a hybrid of hierarchical divisive and k-means clustering) (Zhao & Karypis, 2003). The specific formulation of each system is shown here:

- PB1: first order target co-occurrence features, average link clustering in similarity space.
- PB2: first order target co-occurrence features, repeated bisections in vector space.
- PB3: first order bigram features, average link clustering in similarity space.

Purandare and Pedersen (2004) report that the PB methods generally performed better where there was a reasonably large amount of data available (i.e., several thousand contexts).

Second Order Methods

The second order methods are based on Schütze (1998) and are referred to as SC1, SC2, and SC3. They rely on bigram and co-occurrence features. However, rather than identifying which of these features occur in a context to be clustered, an indirect second order representation is created. The bigram or co-occurrence features are the basis of matrices, where each row and column represents a first order vector for a given word.

The bigram matrices are asymmetric, where the rows represent the first words of the bigrams, and the columns represent the second words. Every cell $[i,j]$ contains the log-likelihood ratio of the bigram formed by the i^{th} row-word followed by the j^{th} column-word.

The co-occurrence matrices are symmetric, where the rows and columns represent the same set of words. Again, the cell values indicate the log-likelihood ratio of the co-occurrences formed by the corresponding row and column words.

The matrices are then (optionally) reduced by Singular Value Decomposition (SVD) to retain the minimum of 300 and 10% of the number of columns, thereby reducing the dimensionality of the feature space.

Like the PB methods, the SC methods are clustered in similarity or vector space using average link agglomerative clustering or repeated bisections. Their configurations consist of:

- SC1: second order co-occurrence features, repeated bisections in vector space.
- SC2: second order co-occurrence features, average link clustering in similarity space.
- SC3: second order bigram features, repeated bisections in vector space.

Purandare and Pedersen (2004) found that SC methods fare better than PB methods when clustering smaller amounts of data (i.e., 50-200 instances) and in capturing fine sense granularities as exhibited by the SENSEVAL-2 corpus.

4 Experimental Data

Our experimental data is the Word Sense Disambiguation Set³ from the NLM. This data is manually tagged with senses drawn from the UMLS. It is important to understand that the UMLS

³ To obtain the WSD set from the NLM's web site, a user needs to register for a free UMLS license: <http://umlsks.nlm.nih.gov/kss/servlet/Turbine/template>

is significantly different than a dictionary, which is often the source of the sense inventory used in manual sense tagged. Rather, the UMLS integrates more than 100 medical domain controlled vocabularies such as SNOMED-CT and the International Classification of Diseases (ICD). UMLS has three main components⁴. The Metathesaurus includes all terms from the controlled vocabularies and is organized by concept, which is a cluster of terms representing the same meaning. The Semantic Network groups the concepts into 134 types of categories and indicates the relationships between them. The Semantic Network is a coarse ontology of the concepts. The SPECIALIST lexicon contains syntactic information for the Metathesaurus terms. Medline is the NLM's premier bibliographic database which includes approximately 13 million references to journal articles in life sciences with a concentration on biomedicine⁵.

In this study, we work with two training sets. The small training set is the NLM WSD set which comprises 5000 disambiguated instances for 50 highly frequent ambiguous UMLS Metathesaurus strings (Weeber et al., 2001). Each ambiguity has 100 manually sense-tagged instances. All instances are derived from Medline abstracts. Twenty one of the ambiguities have fairly balanced sense distributions (45-79% majority sense), while the remaining 29 have more skewed distributions (80-100% majority sense). Each ambiguity is provided with the sentence it occurred in and also the Medline abstract text it was derived from. Every ambiguity has a "none of the above" category which captures all instances not fitting the available UMLS senses, but does not necessarily represent a monolithic sense. Table 1 presents the NLM WSD words and their UMLS senses. Full description of each ambiguity, its senses and UMLS mappings can be found on http://wsd.nlm.nih.gov/Restricted/Reviewed_Results/index.shtml and in [1].

The second training set, or the large training set, is a reconstruction of 1999 Medline which was used in (Weeber et al., 2001). We identified all forms of the NLM WSD set ambiguities occurring in that set and matched them against the 1999 Medline abstracts. The matched abstracts were then used to create the large training set instances. Column 1 in Table 3 lists the training instances for each word. It must be noted that our counts differ slightly from the ones reported in (Weeber et al., 2001). The main reason is that we excluded matches in the titles and restricted the search to the forms occurring in the NLM WSD set regardless of whether Metamap could provide single or multiple mappings.

Experiment 1 describes results on the small set only. Experiment 2 reports results on both small and large sets.

⁴ <http://www.nlm.nih.gov/research/umls/>

⁵ <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

Table 1: NLM WSD set – words, sense definitions and number of instances per sense

Word/Ambiguity	Sense definition	N	Word/Ambiguity	Sense definition	N	Word/Ambiguity	Sense definition	N
adjustment	Individual Adjustment	18	fluid	Liquid substance	100	reduction	Reduction - action	2
	Adjustment Action	62		Fluid behaviour	0		Reduction (chemical)	9
	Psychological Adjustment	13		none of the above	0		none of the above	89
association	none of the above	7	frequency	Frequencies	94	repair	Repair - action	52
	Mental association	0		Increased frequency of micturition	0		Wound healing	16
	Relationship by association	0		none of the above	6		none of the above	32
blood pressure	none of the above	100	growth	Organism function	37	resistance	Psychotherapeutic	3
	Blood pressure	54		Functional concept	63		Social resistance	0
	Blood pressure determination	2		none of the above	0		none of the above	97
cold	Arterial pressure	44	immunosuppression	Therapeutic immunosuppression	59	scale	Integumentary scale	0
	none of the above	0		Natural immunosuppression	41		Intellectual scale	65
	Cold temperature	86		none of the above	0		Weight measurement scales	0
culture	Common cold	6	implantation	Blastocyst implantation, natural	17	secretion	none of the above	35
	Chronic Obstructive Airway Disease	1		Implantation procedure	81		Bodily secretion	1
	Cold therapy	0		none of the above	2		process of secretion	99
condition	Cold sensation	2	inhibition	Psychological inhibition	1	sensitivity	none of the above	0
	none of the above	5		Physical inhibition	98		Statistical sensitivity	49
	Condition	90		none of the above	1		Personality sensitivity	1
degree	Conditioning (Psychology)	2	japanese	japanese language	6	sex	Antimicrobial susceptibility	1
	none of the above	8		japanese population	73		none of the above	49
	Anthropological culture	11		none of the above	21		Coitus	15
depression	Laboratory culture	89	ganglion	Benign cystic mucinous tumor	7	single	Individual behaviour	5
	none of the above	0		Ganglia	93		Gender	80
	Qualitative concept	63		none of the above	0		none of the above	0
determination	Intellectual product	2	glucose	Substance	91	strains	unmarried	1
	none of the above	35		Glucose measurement	9		singular	99
	Mental depression	85		none of the above	0		none of the above	0
discharge	Depression motion	0	lead	lead (element)	27	support	Muscle strain	1
	none of the above	15		Lead measurement	2		Microbiology subtype strains	92
	adjudication	0		none of the above	71		none of the above	7
energy	determination aspects (lab procedure)	79	man	male (organism attribute)	58	surgery	support, device	8
	none of the above	21		Population group (men)	1		supportive care	2
	Discharge, Body substance	1		Homo sapiens	33		none of the above	90
evaluation	Patient discharge	74	mole	none of the above	8	transient	Surgery specialty	2
	none of the above	25		mol (quantitative concept)	83		Surgery	98
	Vitality	1		mole the mammal	1		none of the above	0
extraction	Energy (physics)	99	mosaic	Benign melanocytic nevus of skin	0	transport	Transient	99
	none of the above	0		none of the above	16		Transient population group	1
	Evaluation (functioning concept)	50		Spatial mosaic	45		none of the above	0
failure	Health evaluation	50	nutrition	Embryonic mosaic	52	ultrasound	Biological transport	93
	none of the above	0		none of the above	3		Patient transport	1
	Laboratory procedure	82		nutrition (organism attribute)	45		none of the above	6
fat	Therapeutic procedure	5	pathology	Science of nutrition	16	variation	Ultrasonography	84
	none of the above	13		Feeding and dietary regimes	28		Ultrasonic shockwave	16
	Biological function	4		none of the above	11		none of the above	0
fit	Personal failure	25	pressure	Occupation or discipline	14	weight	Variation (Genetics)	20
	none of the above	71		Pathologic function	85		Variant	80
	Obese build	2		none of the above	1		none of the above	0
radiation	Fatty acid glycerol esters	71	radiation	Pressure - physical agent	96	white	Qualitative concept	24
	none of the above	27		Pressure - action	0		Body weight	29
	Seizures	0		Baresthesia	0		none of the above	47
radiation	Fit and well	18	radiation	none of the above	4	white	white color	41
	none of the above	82		Electromagnetic energy	61		Caucasoid race	49
				Radiation therapy	37		none of the above	10
			none of the above	2				

5 Evaluation

We evaluate the efficacy of the clustering algorithms by determining the mapping from the discovered clusters to the true sense tags that result in maximal accuracy. For all experiments, our test set is the NLM WSD one; it is the training corpus that differed. Experiment 1 deals with only with the small set; experiment 2 uses both the small and the large sets.

For the small training set, the 100 contexts associated with each target word are each treated as a corpus, and features are identified from those contexts for use during clustering. Note that all 100 instances were used for training, clustering and evaluating/testing the clusters, which is not uncommon for unsupervised methods. The window size is set to 5. For the large training set, features are extracted from all instances associated with the target ambiguity. The window size is set to 2.

We report our results in terms of the F-score which is the harmonic mean of the precision and recall. Precision is the number of correctly clustered instances divided by the number of clus-

tered instances; recall is the number of correctly clustered instances divided by all instances. There may be some number of contexts that the clustering algorithm declines to process, which leads to the difference in precision and recall. Our baseline is a simple clustering algorithm that assigns all instances of a target word to a single cluster. The precision and recall of this approach is equivalent to the distribution of the majority sense, which is the percentage associated with the predominant sense of the target word.

The number of clusters to be found must be specified. We set the number to exact number of senses which is equal to the senses assigned by the UMLS plus a “none of the above” category.

The reported statistical results use a t-test for paired two sample means and a level of significance of 0.05. The null hypothesis is that there is no difference.

6 Experimental Results – Experiment 1

We ran experiments for the three SC and the three PB configurations. For the three SC configurations, we ran two sets of experiments, both with and without SVD. We ran all experiments using the sentence and then the abstract as the context of the target word. All experiments were run seeking six clusters and then the exact number of senses as found in the manually tagged data. The choice of six clusters is based on the fact that this is more than the maximum number of possible senses for any word observed in this data (most words have 2-3 senses). We believe that an effective clustering method should identify approximately the correct number of clusters and leave any “extra” clusters relatively unpopulated. We cluster with the exact number of senses to test this hypothesis.

Table 2 summarizes the results. The words are grouped according to the majority sense. For each word, the best method from our PB and SC experimental configurations for six and the exact number of clusters is listed along with its F-score (columns 3-6). F-scores equal or greater than the majority sense are bolded.

When finding *six clusters*, our best methods perform above the majority sense for 16 out of 21 words for 45-79% majority sense and 6 out of 29 words for the skewed sense distribution of 80-100% majority sense. For 45-79% majority sense, our best methods with F-scores in Table 2, column 4 are significantly better than the baseline (p-value<0.05); for 80-100% with F-scores in Table 2, column 4 the baseline performs significantly better (p-value<0.05). When all results from the best methods for six clusters are considered, there is no significant difference between best methods and baseline (p-value>0.05).

For the *exact number of clusters*, our best methods are above the majority sense for 20 out of 21 ambiguities with 45-79% majority sense, and 12 out of 29 words for 80-100% majority sense. For 45-79% majority sense, our best methods with F-scores in Table 2, column 6 are significantly better than the baseline (p-value<0.05); for 80-100% majority sense with F-scores in Table 2, column 6 there is no significant difference (p-value>0.05). When all results from the best methods for exact clusters are considered, our best methods perform significantly better than the baseline (p-value<0.05). F-scores from best methods for exact clusters (Table 2, column 6) are significantly better than the F-scores for six clusters (Table 2, column 4) for 45-79%, 80-100% and all sense distributions (p-value<0.05).

SC2 configurations with or without SVD are consistently the top methods. Out of 56 paired comparisons, SC2 methods are significantly better in 47 cases (p-values<0.05) and not significantly different in 7 (p-values>0.05 for SC2_noSVD_a v. PB3_a, SC2_SVD_s v PB1_s, SC2-

SVD_s v PB3_s, SC2_SVD_s v SC3_SVD_s, SC2_noSVD_a v PB3_a, SC2_SVD_s v PB1_s, SC2_SVD_s v PB3_s, SC2_SVD_s v SC1_SVD_s and SC2_SVD_a v PB3_a).

Overall the abstract as context provides better discrimination results than does the sentence context (p-values<0.05 except for SC2_s v SC2_a, where p-value>0.05).

The application of SVD to the matrix divided the methods performance into three categories. The methods positively influenced by the application of SVD are SC1 and SC3 with 6 clusters and context=sentence or context=abstract, SC1 and SC3 with exact number of clusters and context=sentence (p-values<0.05). There is no significant difference with SVD on or off for SC2 with 6 clusters and context=abstract, and SC1, SC2 and SC3 with exact clusters and context=abstract (p-values>0.05). Two methods are negatively influenced by SVD – SC2 with exact clusters and context=sentence and 6 clusters and context=sentence (p-value<0.05).

Table 2: Experiment 1 - summary of results (sorted by Majority sense; F-scores from experimental methods equal or greater than Majority sense are bolded; -SVD indicates application of SVD; -a indicates context=abstract; -s indicates context=sentence)

Word/ambiguity	Majority sense (in %)	6 clusters best result: method	6 clusters best result: F-score	Clusters = exact number of senses best result: method	Clusters = exact number of senses best result: F-score
nutrition	45.00	PB1-a; SC2-a-SVD	45.92	PB3-a	47.00
weight	47.00	SC3-a	57.32	PB2-a	61.71
sensitivity	49.00	PB3-s	54.54	PB3-s	55.00
white	49.00	PB3-s	53.12	SC1-a-SVD	57.00
evaluation	50.00	SC1-a-SVD	58.48	SC1-a	62.07
mosaic	52.00	SC2-a	56.12	PB3-s	54.64
repair	52.00	SC1-a	62.50	SC3-a	79.40
blood pressure	54.00	SC3-s-SVD	62.77	SC3-s-SVD	60.92
man	58.00	SC3-a-SVD	63.92	SC1-a-SVD; SC3-a-SVD	65.00
immunosuppression	59.00	SC2-a	59.49	PB2-a	60.92
radiation	61.00	SC2-s	62.95	SC2-s	64.00
adjustment	62.00	PB1-a	66.67	SC2-a	65.00
degree	63.00	SC2-s-SVD	71.79	sc1-s-SVD	71.00
growth	63.00	PB3-a	62.24	PB1-a	64.32
scale	65.00	SC3-s-SVD	63.92	SC2-s; PB3-a; SC2-a/SVD	64.29
failure	71.00	PB1-a	73.10	SC2-a	74.00
fat	71.00	SC2-a	71.43	PB1-s	75.00
lead	71.00	SC2-s	70.05	SC3-a-SVD	79.00
japanese	73.00	SC2-s	72.08	SC2-a	74.00
discharge	74.00	SC2-s; SC2-a-SVD	74.11	PB1-s	76.00
determination	79.00	SC2-a-SVD	78.57	SC2-s	80.40
average 45-79% majority sense	60.38		63.86		66.22
sex	80.00	SC2-a-SVD	78.17	PB3-a; SC2-a-SVD	77.39
variation	80.00	SC2-a-SVD	78.57	SC2-s; PB3-a; SC2-a/SVD	78.39
implantation	81.00	SC2-a	79.19	SC2-s; SC2-a/SVD	81.00
extraction	82.00	PB3-a; SC3-a-SVD	81.22	PB3-a; SC2-a-SVD	83.00
fit	82.00	SC2-a	82.05	sc2-s; SC2-a-SVD	82.41
mole	83.00	SC2-s	83.25	SC2-s-SVD	89.90
ultrasound	84.00	PB3-a	85.13	SC2-a-SVD	84.42
depression	85.00	SC2-a-SVD	87.75	SC2-a-SVD	87.44
pathology	85.00	PB3-a; SC2-a/SVD	83.25	SC2-s; PB3-a; SC2-a/SVD	83.00
cold	86.00	SC2-a	86.74	SC2-a	86.74
culture	89.00	SC2-s/SVD; PB3-a; SC2-a-SVD	85.71	SC2-s/SVD; PB3-a; SC2-a/SVD	87.44
reduction	89.00	PB3-a; SC2-a	85.28	SC2-s; PB3-a; SC2-a/SVD	87.00
condition	90.00	SC2-a-SVD	91.37	SC2-a-SVD	91.00
support	90.00	SC2-s; PB3-a; SC2-a/SVD	86.29	SC2-s; PB3-a; SC2-a/SVD	88.00
glucose	91.00	SC2-a-SVD	86.91	SC2-a/SVD	91.46
strains	92.00	SC2-a-SVD	91.37	SC2-a-SVD	92.00
ganglion	93.00	PB1-a	90.82	SC3-s-SVD	92.46
transport	93.00	PB1-a	91.84	PB1-a	92.00
frequency	94.00	PB3-a	92.86	PB1-a	95.48
pressure	96.00	SC2-a/SVD	92.86	SC2-s; SC2-a/SVD; SC1-s-SVD; SC3-s-SVD	93.94
resistance	97.00	PB3-a; SC2-a	93.88	SC2-s; PB3-a	95.48
inhibition	98.00	SC2-s	94.84	SC2-s	96.00
surgery	98.00	PB3-a; SC2-a-SVD	94.36	SC2-s; PB3-a; SC2-a/SVD	96.48
energy	99.00	SC2-s; PB3-a; SC2-a/SVD	95.92	SC2-s-SVD; PB3-a; SC2-a/SVD	97.49
secretion	99.00	SC2-a-SVD	95.92	PB3-a; SC2-a-SVD	97.49
single	99.00	SC2-s; PB3-a; SC2-a/SVD	95.92	SC2-s; PB3-a; SC2-a/SVD	97.49
transient	99.00	SC2-a	95.38	SC2-s; SC1-s-SVD	97.49
association	100.00	SC2-a/SVD	97.44	PB3-a; SC2-a/SVD	98.99
fluid	100.00	PB1-a; SC2-a/SVD; SC3-s-SVD	97.44	SC2-s/SVD; PB3-a; SC2-a/SVD	98.99
average 80-100% majority sense	90.83		89.03		90.43
average all	78.04		78.46		80.26

The scope of contexts in Purandare and Pedersen (2004) was limited to 2-3 sentences, and there were approximately 100 contexts to cluster. The NLM WSD data also consists of 100 contexts per target word, so we initially hypothesized that our results would support their conclusion that clustering contexts represented as second order feature vectors using the method of repeated bisections in vector space would give the best results (SC1 or SC3).

However, our findings differ in various ways. Our most successful method is SC2, which uses second order contexts and agglomerative clustering in similarity space. We also noticed that PB1 and PB3, both of which use agglomerative clustering in similarity space, significantly outperform PB2 (p -values <0.05). Thus, rather than using vector spaces, our results suggest the use of similarity space.

Not surprisingly, using the entire abstract as the context leads to overall better results than single sentences. This is consistent with Liu, Teller and Friedman (2004) who find that their supervised classifiers for the biomedical domain yield better results when a paragraph of context is used as opposed to a 4-10 word window for general English classifiers.

The larger scope of context provided by an abstract gives us a rich collection of features which compensates for the smaller overall number of contexts we observe in the NLM WSD data. In the case of the SENSEVAL-2 data as used by Purandare and Pedersen, the scope of the individual contexts was small, as was the number of contexts. This led to a smaller feature space where it was impossible to find meaningful pair-wise similarities among the contexts. In the NLM WSD data however, individual context vectors are represented in high dimensional feature spaces and are rich enough to allow for agglomerative clustering in pair-wise fashion.

However, if the NLM WSD data provides many features per instance, why don't our results show better performance of first order methods, as Purandare and Pedersen would predict? We would argue that our medical journal abstract text is much more domain-specific and has a more restricted vocabulary. As such our contexts are more focused than general English text. The restricted nature of our corpus introduces less noise in the second order representations, and allows them to perform better than first order representations which generally require larger number of instances to provide enough features to perform well (Purandare and Pedersen, 2004).

The application of SVD produced mixed results. We attribute the overall lack of improvement demonstrated by SVD to the fact that we are creating second order vectors by averaging the word vectors for all the words in our context, which is the entire abstract. This means that our averaged vector is based on a large number of word vectors, and there may be a considerable amount of noise in the resulting averaged vector. We are currently exploring methods of selecting the word vectors to be used for building the second order representation in a more restrained fashion.

As it was pointed out in Section 4, the "none of the above" category does not necessarily represent a monolithic sense. It is possible that our methods subdivide this category into finer groups whose correctness needs to be determined by additional human expert evaluation which we did not perform for this study. These finer groups could potentially become newly discovered senses to be included in the ontological tree.

7 Experimental Results – Experiment II

For the large and small training set, for both the PB and SC configurations we used the entire abstracts as our contexts. We ran each SC and PB configuration with and without SVD.

Table 3 summarizes the results. The words are grouped by majority sense. For each word, the best method from our PB and SC experimental configurations is listed along with its F-score

(columns 3-6). Columns 3-4 are results from small training set; columns 5-6 are those from the large training set.

Our best methods are above the majority sense for 20 out of 21 ambiguities for the small training set and 19 out of 21 words for the large training set with 45-79% majority sense, and 10 out of 29 words for the small training set and 9 out of 29 words for the large training set for 80-100% majority sense. SC2 configurations with or without SVD are consistently the top methods on both the small and large training sets.

We aimed to evaluate the contributions of more data to the performance of the proposed methods. For four methods, it did not have a significant effect (PB1_noSVD, PB3_noSVD, SC1_noSVD, SC3_noSVD with p -values >0.05). For three methods, it had a significant positive effect (PB2_noSVD, SC1_SVD, SC3_SVD with p -values <0.05). Surprisingly, for two of the methods (SC2_noSVD and SC2_SVD), more data had a significant negative effect (p -values <0.05) which means that more data lowered the average F-scores. These results suggest that our methods work well on both large datasets and small sets with good representations of the senses.

Our best performing methods, SC2, on average perform worse when trained on more data. A possible explanation is the unique combination of features (second order co-occurrences) and clustering method (average link agglomerative clustering in similarity space). Those features create a rich enough representation from the small training set for meaningful pair-wise similarity aggregations. On the other hand, if the same clustering method is used but with first order features, on the average more training data does not influence the results. This points to the stability of agglomerative clustering methods with second order features extracted from small sets. Repeated bisection clustering performs better with features extracted from larger training sets as demonstrated by the performance of PB2_noSVD, SC1_SVD and SC3_noSVD configurations.

Another goal of the experiments was to evaluate the performance of the methods with and without SVD. The contributions of SVD are not pronounced in both our small and large training sets (p -values >0.05 for 7 out of 12 pairs). The only methods influenced positively by SVD are PB3 and SC3 with the large training set (p -values <0.05). We attribute the overall lack of improvement demonstrated by SVD to the fact that we are creating second order vectors by averaging the word vectors for all the words in our context, which is the entire abstract. This means that our averaged vector is based on a large number of word vectors, and there may be a considerable amount of noise in the resulting averaged vector. We are currently exploring methods of selecting the word vectors to be used for building the second order representation in a more restrained fashion.

For first order features in particular, the larger scope of context provided by an abstract gives us a rich collection of features which compensates for the smaller overall number of contexts in the small training set. Individual context vectors are represented in high dimensional feature spaces and are rich enough to allow for agglomerative clustering in pair-wise fashion employed in our best performing methods (SC2).

Table 3: Experiment 2 – Summary of results (sorted by Majority sense; F-scores from experimental methods equal or greater than Majority sense are bolded; context=abstract; -SVD suffix indicates application of SVD)

<i>Word/ambiguity (training instances from small training set/training instances from large training set)</i>	<i>Majority sense</i>	<i>Best result: method (small training set: NLM WSD set)</i>	<i>Best result: F-score (small training set)</i>	<i>Best result: method (large training set: 1999 Medline)</i>	<i>Best result: F-score (large training set)</i>
nutrition (100/2637)	45%	PB3	47.00	PB1, PB1-SVD, SC2-SVD	45.00
weight (100/14121)	47%	PB2	61.71	PB2	69.00
sensitivity (100/21304)	49%	PB1	51.00	PB1	70.35
white (100/4467)	49%	SC1-SVD	57.00	SC1-SVD	60.00
evaluation (100/12453)	50%	SC1	62.07	SC1-SVD	68.72
mosaic (100/402)	52%	SC2	53.27	SC2-SVD	58.58
repair (100/3960)	52%	SC3	79.40	PB2, PB2-SVD	78.39
blood pressure (100/4767)	54%	PB3	55.28	SC2	53.53
man (100/3852)	58%	SC1-SVD; SC3-SVD	65.00	SC2	77.00
immunosuppression (100/956)	59%	PB2	60.92	SC2	73.57
radiation (100/4386)	61%	PB1; PB3; SC2-SVD	61.00	SC3-SVD	66.00
adjustment (100/1772)	62%	SC2	65.00	PB3-SVD	63.00
degree (100/17694)	63%	PB2	68.89	SC3-SVD	66.00
growth (100/20794)	63%	PB1	64.32	PB2	65.91
scale (100/5719)	65%	PB3; SC2, SC2-svd	64.29	SC2-SVD	67.35
failure (100/10049)	71%	SC2	74.00	PB3	72.00
fat (100/5940)	71%	SC2, SC2-SVD	71.00	SC2	73.00
lead (100/7088)	71%	SC3-SVD	79.00	SC2-SVD	69.00
japanese (100/1662)	73%	SC2	74.00	PB1, PB1-SVD	79.00
discharge (100/3056)	74%	PB1; PB3; SC2, SC2-SVD	74.00	PB3	76.00
determination (100/28256)	79%	PB3; SC2	79.40	SC2-SVD	79.40
average 45-79% majority sense	60%		65.12		68.13
sex (100/6522)	80%	PB3; SC2-SVD	77.39	PB3	83.42
variation (100/7781)	80%	PB3; SC2, SC2-SVD	78.39	PB3	79.40
implantation (100/2302)	81%	SC2, SC2-SVD	81.00	PB3	79.00
extraction (100/7092)	82%	PB3; SC2-SVD	83.00	SC2-SVD	83.00
fit (100/2426)	82%	PB3; SC2, SC2-SVD	82.41	PB1, PB1-SVD	81.41
mole (100/1797)	83%	SC2, SC2-SVD	82.41	PB1, SC2-SVD	81.82
ultrasound (100/3082)	84%	SC2-SVD	84.42	SC2-SVD	84.42
depression (100/4335)	85%	SC2-SVD	87.44	SC2-SVD	85.28
pathology (100/3217)	85%	PB3; SC2, SC2-SVD	83.00	PB1, PB1-SVD; SC2-SVD	66.00
cold (100/1417)	86%	SC2	86.74	PB1, PB1-SVD	84.26
culture (100/15398)	89%	PB3; SC2, SC2-SVD	87.44	PB3	87.44
reduction (100/15853)	89%	PB3; SC2, SC2-SVD	87.00	PB1-SVD	89.00
condition (100/19384)	90%	SC2-SVD	91.00	PB1	88.00
support (100/16849)	90%	PB3; SC2, SC2-SVD	88.00	PB3-SVD	91.00
glucose (100/5896)	91%	SC2, SC2-SVD	91.46	PB3	89.80
strains (100/6741)	92%	SC2-SVD	92.00	SC2-SVD	90.00
ganglion (100/1184)	93%	PB1; PB3; SC2-SVD	92.46	PB1, PB1-SVD	92.46
transport (100/5527)	93%	PB1	92.00	PB3	96.00
frequency (100/12517)	94%	PB1	95.48	SC2-SVD	92.46
pressure (100/8410)	96%	SC2, SC2-SVD	93.94	PB3	95.92
resistance (100/8814)	97%	PB3	95.48	PB3	96.48
inhibition (100/13557)	98%	SC2	96.00	PB1	96.00
surgery (100/14767)	98%	PB3; SC2, SC2-SVD	96.48	SC2	97.46
energy (100/5723)	99%	PB3; SC2, SC2-SVD	97.49	SC2-SVD	99.50
secretion (100/7228)	99%	PB3; SC2-SVD	97.49	PB1, PB1-SVD; PB3-SVD; SC2-SVD	97.49
single (100/22447)	99%	PB3; SC2, SC2-SVD	97.49	PB1-SVD	97.49
transient (100/5526)	99%	SC2, SC2-SVD	96.97	PB1	96.97
association (100/12182)	100%	PB3; SC2, SC2-SVD	98.99	SC2; SC2-SVD	98.48
fluid (100/6759)	100%	PB3; SC2, SC2-SVD	98.99	PB1, SC2, SC2-SVD	98.99
average 80-100% majority sense	91%		90.08		89.62
average all	78%		79.60		80.60

8 Future Work

In this paper there was no stemming or normalization of the text carried out. We plan to conduct experiments where we stem the data, so as to hopefully reduce the scarcity in the feature vectors. We will use the Lexical Variant Generator⁶, a text normalization tool provided by NLM. Currently the clusters that are discovered are not labeled with any sense or definition information. We are now exploring the use of collocation discovery techniques to analyze the text that makes

⁶ <http://med-info.com/lvg.html>

up a cluster to generate a simple label, and will continue to extend that approach in the hopes of arriving at an approximation of a definition or gloss for each cluster. We are also actively working on automatic cluster stopping.

9 Conclusions

This paper shows that methods of unsupervised word sense disambiguation created for the general English domain are indeed suitable for the biomedical domain. In particular, methods based on second order representations of entire abstracts in which a target word appears are more effective in resolving ambiguity, and that these methods are particularly successful when contexts are clustered in similarity space using agglomerative clustering.

Acknowledgements

The research was partially supported by grant NLM 07453-01 and a NSF Faculty Early Career Development (CAREER) award (#0092784). We are very grateful to Jim Mork and Dr. Alan Aronson from the NLM for their unwavering assistance. This work was carried out in part using hardware and software provided by the University of Minnesota Supercomputing Institute.

References

- Chen, L.; Hongfang, L. and Friedman, C. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*. Vol. 21. no 2 2005, pp.248-256.
- Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*. Volume 19, no 1, pp. 61-74.
- Friedman, C. 2000. A broad coverage natural language processing system. 2000. Proc. AMIA. Philadelphia, PA: Hanley and Belfus. pp. 270-274.
- Liu, H.; Lussier, Y. and Friedman, C. 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *JBIM* 34, 249-261.
- Liu, H.; Teller, V. and Friedman, C. 2004. A multi-aspect comparison study of supervised word sense disambiguation. *JAMIA*, vol. 11, no 4, June/August 2004, pp. 320-331.
- Mihalcea, R. 2003. The role of non-ambiguous words in natural language disambiguation. RANLP-2003, Borovetz, Bulgaria.
- Pantel, P.; Lin, D. 2002. Discovering Word Senses from Text. Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
- Pedersen, T. and Bruce R. 1997. Distinguishing word senses in untagged text. Proc. EMNLP. Providence, RI
- Purandare, A. and Pedersen, T. 2004. Word Sense Discrimination by clustering similar contexts. Proc. CoNLL. Boston, MA.
- Roth L. and Hole WT. 2000. Managing name ambiguity in the UMLS Metathesaurus. Proc. AMIA.
- Schütze, H. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, vol. 24, number 1.
- Weeber, M.; Mork, J. and Aronson, A. 2001. Developing a test collection for biomedical word sense disambiguation. Proc. AMIA.
- Widdows, D.; Peters, S.; Cederberg, S.; Steffen, D. and Buitelaar, P. 2003. Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. Workshop on NLP in Biomedicine, ACL, pp. 9-16, Sapporo, Japan.
- Zhao, Y.; Karypis, G. 2003. Hierarchical Clustering Algorithms for Document Datasets. Tech. report 03--027, U of Minnesota, Dept. of Computer Science.