National Library of Medicine Participation Program Report

Bridget T. McInnes
Computer Science Department
University of Minnesota Twin Cities
Minneapolis, MN 55155, USA
bthomson@cs.umn.edu

Abstract

This report introduces the supervised and unsupervised approaches to disambiguate words in biomedical text that have been explored while participating in the National Library of Medicine's Research Participation Program. We explored using biomedical information extracted from the Unified Medical Language System (UMLS) as features in our supervised and unsupervised approaches. Historically supervised approaches to word sense disambiguation have obtained a higher accuracy than unsupervised. We show that this is slowly changing. The unsupervised approaches are obtaining disambiguation accuracies that are comparable to that of supervised approaches.

1 Introduction

Some words have multiple senses, for example, the word $culture$ could mean an anthropological culture, such as the culture of a Mayan civilization, or a laboratory culture, such as a cell culture. The different senses of a word are often obtained from a sense inventory such as a dictionary or other resource. The Unified Medical Language System (UMLS) is one such sense inventory for the biomedical and clinical domain. In the UMLS, senses (or concepts) associated with words and terms are enumerated via Concept Unique Identifiers (CUIs). For example, the two senses of $culture$ are C0010453 (Anthropological Culture) and C0430400 (Laboratory Culture) in the UMLS release 2007AB.

Word sense disambiguation is the task of identifying the appropriate sense of a word that has multiple senses. For example, in the sentence "The culture count doubled", $culture$ as previously stated as two possible senses as stated above. The goal is to automatically identify the appropriate concept given the context the word is used.

There are two main approaches that have been used in word sense disambiguation (WSD): supervised and unsupervised approaches. Supervised approaches use manually annotated training data to build classifiers to determine the appropriate sense of the word. Unsupervised approaches do not require manually annotated data. They use, for example, distributional characteristics of an unannotated corpora or an external knowledge source. The advantage of using supervised approaches is that historically they obtain a much higher disambiguation accuracy than unsupervised. The problem with supervised approaches though is that they require data for each word that needs to be disambiguated and this data is manually created making them unscalable for large scale systems.

In this report, we first provide some background information on the UMLS and its resources. Second, we describe the work that we have done in supervised and then unsupervised WSD. Third, we briefly discuss the state of supervised and unsupervised approaches in the biomedical domain. Lastly, we discuss our future plans.

2 Background

The UMLS is more than just a dictionary of different word senses but also a framework encoded with different semantic and syntactic structures. Some such information includes related concepts,

semantic types and semantic relations. A semantic type is a broad subject categorization assigned to a CUI. A semantic relation is the relationship between two semantic types. For example, the semantic type of C0010453 (Anthropological Culture) is "Idea or Concept" while the semantic type for C0430400 (Laboratory Culture) is "Laboratory Procedure". The semantic relations between "Idea or Concept" and "Laboratory Procedure" with the semantic type "Mental Process" are "result_of" and "assesses_effect_of", respectively. Currently, there exists approximately 1.5 million CUIs and 135 STs in the UMLS.

Medline is an online database that contains 11 million biomedical citations. MetaMap (Aronson, 2001) is a concept mapping system that maps terms in biomedical text such as Medline abstracts to concepts in the UMLS by identifying the CUIs of the content words in the text. It was developed to improve retrieval of biomedical articles such as MEDLINE citations. MetaMap has five components: the preprocessor, the lexical variant generation (LVG) module, the candidate retrieval module, the candidate evaluation module and the mapping construction module.

The preprocessor has three steps: i) the terms in the input data are identified using the SPECIALIST Lexicon, ii) the input data is part-of-speech tagged using the Xerox POS tagger, and iii) the input data is parsed using the SPECIALIST minimal commitment parser. The LVG module generates variants for each term in the input data using the SPECIALIST Lexicon. The candidate retrieval module, identifies potential concepts from the Metathesaurus for each term in the input data. A potential concept is chosen because it contains at least one of the variants in its string. For example: "Vena Cava Filter" and "Stents" would both be possible concepts for the term "inferior vena cava stent filter". The candidate evaluation module assigns a MetaMap score to each concept based on four criteria: centrality, variation, coverage and cohesiveness. Centrality is whether the potential concept contains the head of the input data term. Variation is the distance between the input data term and potential Concept. Coverage is the length of the term versus the concept. For example, the term "inferior vena cava stent filter" contains five words while the possible concepts "Vena Cava Filter" and "Stents" respectively contain three and one. Cohesiveness is how continuous the match between the term and the concept is. For example, for the term "inferior vena cava stent filter" and the potential concept "Vena Cava Filter" have two words the consecutively overlap. The concept with the highest MetaMap score is assigned to the associated term. The mapping construction module generates the "Concept Mapped Medline Abstract".

MetaMap is used by the Medical Text Indexer (MTI) which is an indexing system that recommends Medical Subject Headings (MeSH) about an abstracts to medical text indexers. The medical text indexer assigns medical abstracts with MeSH headings for indexing purposes. MeSH headings are concepts that are obtained from the MeSH vocabulary. MeSH is a controlled vocabulary specifically designed for the indexing of biomedical articles. In MTI, the possible MeSH headings of an abstract are ranked partially based on their MetaMap Indexing (MMI) score which is the product of a frequency factor and relevance factor. The frequency factor takes into account how many times the CUI is mapped to a term in the text. The relevance factor is a weighted average of four components: MeSH tree depth, word length, character count and MetaMap score.

3 Supervised WSD

There has been previous work on supervised learning approaches to word sense disambiguation in the biomedical domain. However, some of that work has simply used features that are known to work in general English, and do not take advantage of biomedical information (e.g., (Liu et al., 2004), (Joshi et al., 2005)).

On the other hand, Leroy and Rindflesch (Leroy and Rindflesch, 2005) include the use of biomedical information generated by MetaMap. They analyze combinations of the following features: i) whether the target word is a head word, ii) the part-of-speech (POS) of the target word, iii) the semantic relations between the words surrounding the target word and between surrounding words and the target word itself, and iv) the semantic types of the words or terms surrounding the target word. Leroy and Rindflesch's best reported feature set contains whether the target word is a head word, the POS of

the target word and the semantic types of the words in the same sentence as the target word. They use the Naive Bayes algorithm from the WEKA (Witten and Frank, 1999) data-mining package and report their results using 10-fold cross validation.

Joshi, et al. (Joshi et al., 2005) employ features that have been used in supervised learning of word sense disambiguation for general English, and apply them to the biomedical domain. Their approach utilizes features based on the unigrams and bigrams of the words in the same window of context as the target word. A unigram is a content word that frequently occurs in a window of context around the target word. A bigram is an ordered pair of content words that frequently occur in a window of context around the target word. Joshi et al. report highly accurate results, especially when their features are unigrams where the window is the same sentence as the target word and unigrams where the window of context is the same abstract as the target word. They compare the Naive Bayesian classifier and Support Vector Machine from the WEKA data-mining package and report their results using 10-fold cross validation.

Liu, et al. (Liu et al., 2004) utilize combinations of the following features: i) surrounding words, ii) orientation, iii) distance, iv) collocations and v) unigrams. Orientation is whether the surrounding word is to the left or the right of the target word. Distance is how far the surrounding word is from the target word and collocation is a unit of words that represent a single idea, for example, "White House". Their best reported feature set contains all word within a window size of three, their orientation, and the three nearest two word collocations. They compare the Naive Bayes, a modified Decision List and a combination Naive Bayes/exemplar-based algorithm. They report their results using 10-fold cross validation and record the best per word accuracy over all feature sets and algorithms.

McInnes, et al. (McInnes et al., 2007) use the biomedical feature Concept Unique Identifiers (CUIs) that frequently occur in a window of context around the target word. McInnes, et al. explore using CUIs in the same sentence as the target word and CUIs in the same abstract as the target word for their window of context. They use the Naive Bayesian classifier from the WEKA data-mining package and report their results using 10-fold cross validation.

Stevenson, et al. (Stevenson et al., 2008) use a combination of biomedical and general English features. The use the general English features collocations, syntactic dependencies and bag-of-words combined with the biomedical feature MeSH headings. They compare the Naive Bayesian classifier, the Support Vector Machine (SVM) and Vector Space Model and report their results using 10-fold cross validation. They found the Vector Space Model significantly outperformed the SVM and Naive Bayes classifier.

In the following sections, we discuss the evaluation data used to evaluate the previous supervised approaches discussed above as well as our own. We then discuss our supervised approach and the results.

4 Supervised Methods

Evaluation Dataset : We use the National Library of Medicine's Word Sense Disambiguation (NLM-WSD) dataset (Weeber et al., 2001). This data contains 100 randomly selected instances of 50 frequent and highly ambiguous words from 1998 MEDLINE abstracts. Each instance of a target word was manually disambiguated by 11 human evaluators who assigned the word a CUI or "None" if none of the CUIs described the concept.

Joshi, et al. evaluated their approach using 28 out of the 50 target words in the dataset; referred to here as Joshi subset. Leroy and Rindflesch evaluated their approach using 15 out of the 28 words used by Joshi, et al.; referred to as Leroy subset. Liu, et. al. evaluated their approach using 22 out of the 28 words; referred to as Liu subset. There are nine words that all three authors use to evaluated their approach; referred to as Common subset. There are 22 words that were not used by any of the authors; referred to as the Excluded subset. These words were not used because a large majority of their instances have the same concept. We report our results for all 50 words as did McInnes, et al. and Stevenson, et. al.

Supervised WSD Approach : We explore using a number of biomedical information as features in a supervised learning approach to word sense disambiguation.

McInnes, et al. use the CUIs that are assigned

to a term as separate features. For example, the term "New York Heart Association" has the following CUIs assigned to it by MetaMap: 1) C0027976 (New York), 2) C1281570 (Heart) and 3) C0699792 (association). McInnes, et al. considers each CUI a separate feature.

McInnes, et al's use of CUIs lead us to explore what we call conflated CUIs.

We experiment with combining the CUIs into what we call a "conflated CUI", treating it as a single feature rather than three separate features.

Joshi, et al. use ngrams as features into their supervised system. An ngram is a sequence of N words that commonly occur together in a text.

Joshi, et al's use of ngrams lead us to explore what we call CUI ngrams.

Instead of words, we use a sequence of N CUIs that commonly occur together in a text as features which we refer to as CUI ngrams. We also experiment with using "semantic type ngrams" which use a sequence of N semantic types as a feature. We experiment with N = 2 and N = 3.

Stevenson, et al. use MeSH concepts associated with the abstract in the NLM-WSD dataset as features into their supervised system. Stevenson, et al's success with using MeSH concepts showed using global context features improved the disambiguation accuracy of supervised systems. The problem with using MeSH descriptors is that they are manually assigned and the automatic indexer, Medical Text Indexer, that provides the indexers with possible MeSH concepts uses MetaMap. This causes us concern because one of the overall goals of this research is to improve MetaMap through WSD. MetaMap ranks CUIs associated with an abstract based on their MMI score. This MMI score is used by the Medical Text Indexer algorithm to determine which MeSH concepts should be provided to the indexer.

Stevenson, et al's use of MeSH terms lead us to explore using the CUIs returned by MetaMap that have an MMI score greater than 10 as features.

Journal Descriptors are terms that are assigned to a journal that describe the type of articles it contains such as "Cardiology", "Surgery", "Health Informatics", and "Amnesia". Journal Descriptor Indexing (JDI) (Humphrey, 1999) is a system that automatically assigns Journal Descriptors to journal titles in order to maintain a subject index of all journals in MEDLINE.

Stevenson, et al's use of global context features lead us to explore using Journal Descriptors as features into our supervised system.

We explored using JDs as assigned by the JDI algorithm as features in our supervised system.

So overall, we have explore different feature sets:

- conflated CUIs
- CUI bigram
- CUI trigram
- semantic type (ST) unigram
- ST bigram
- ST trigram
- MMI
- JDs

We use the Support Vector Machine (SMO) algorithm from the WEKA data-mining package as our learning algorithm and report the accuracy of our approach using 10-fold cross validation. In 10-fold cross validation the instances are divided into ten blocks where each block contains an equal number of instances. Then nine blocks are used as a training data and the remaining block is used as test data. The classifier is built using the nine blocks as training data and tested using the remaining block. This is repeated ten times such that each block has been used as test data exactly once with the other nine as training data. The accuracy reported is the average over all ten runs.

5 Supervised Results and Discussion

Table 1 shows the accuracy (%) of our approaches using the following features into a Support Vector Machine evaluated on the NLM-WSD dataset: i) conflated CUIs, ii) CUI bigram, iii) CUI trigram, iv) ST (semantic type) unigram, iv) ST bigram, v) ST trigram, v) MMI, and vi) JDs. The table also shows the accuracy of the "majority sense" baseline referred to as "baseline" in the table, and the previous approaches introduced by McInnes, et al., and Stevenson, et al. In this section, we compare the accuracy of our different approaches with those reported by McInnes, et al, and Stevenson, et al, as well as the majority sense baseline. The $p\text{-}values$ reported are calculated using the one-sided pairwise t-test.

Table 1: **Accuracy of Supervised Approaches on the NLM-WSD Dataset**

target word	baseline	Stevenson	McInnes	CUI			ST			Global	
		et al.	et al.	conflated	bigram	trigram	unigram	bigram	trigram	JD	MMI
adjustment	62	74	66	64	64	64	71	65	64	65	72
association	54	100	98	97	98	100	98	98	96	100	98
blood pressure	50	46	55	53	48	48	45	54	51	50	57
cold	59	88	89	89	85	78	87	88	87	91	88
condition	61	89	90	90	90	90	85	90	90	88	91
culture	49	95	93	92	89	89	98	95	87	98	92
degree	86	95	78	82	75	69	69	76	74	67	78
depression	85	88	81	80	78	77	78	81	78	94	84
determination	74	87	82	81	79	79	79	81	75	79	87
discharge	82	95	92	94	78	74	88	95	85	84	91
energy	71	98	99	99	96	99	99	99	98	99	98
evaluation	81	81	74	76	65	60	62	73	62	59	73
extraction	73	85	83	83	83	83	83	84	83	87	85
failure	71	67	74	77	71	71	66	72	71	73	72
fat	83	84	77	78	78	73	72	81	80	71	82
fit	85	88	86	83	79	82	78	86	79	85	86
fluid	89	100	99	98	98	100	99	99	98	100	99
frequency	80	94	93	93	92	94	94	93	92	95	93
ganglion	84	96	94	94	94	95	95	94	93	97	95
glucose	63	91	90	90	89	84	87	90	90	91	90
growth	63	68	70	70	68	65	65	72	62	73	65
immunosuppression	58	80	79	70	66	46	65	69	66	78	78
implantation	52	93	94	93	85	85	91	90	87	91	93
inhibition	45	98	98	97	98	98	98	98	98	97	98
japanese	52	75	77	77	78	78	72	80	78	76	77
lead	65	94	88	90	79	75	90	89	87	88	86
man	47	90	86	82	70	62	79	78	75	74	78
mole	49	93	84	85	83	82	83	85	84	76	86
mosaic	100	87	74	77	65	52	78	80	72	85	72
nutrition	90	54	46	49	42	45	34	32	38	45	36
pathology	89	85	83	83	84	85	87	84	84	76	86
pressure	79	95	94	94	95	88	95	95	94	96	94
radiation	99	84	84	83	83	79	71	82	82	78	85
reduction	71	89	89	89	89	89	92	89	89	89	90
repair	82	88	93	92	76	55	91	90	83	88	85
resistance	100	98	96	97	96	97	95	97	93	100	95
scale	94	88	78	81	63	65	79	76	67	75	80
secretion	93	99	99	99	98	99	99	99	97	99	99
sensitivity	91	93	94	91	83	71	86	86	83	72	88
sex	98	87	89	88	82	80	83	82	79	78	87
single	96	99	98	98	98	99	98	98	95	99	99
strains	97	93	92	92	91	92	92	92	92	91	92
support	99	89	90	90	90	90	88	91	90	89	89
surgery	99	97	93	94	94	96	95	95	93	98	98
transient	92	99	99	98	98	99	99	98	97	99	97
transport	90	93	93	94	94	93	92	93	94	89	94
ultrasound	98	90	84	84	83	86	84	84	82	82	86
variation	99	95	87	89	79	80	92	92	79	89	86
weight	93	81	75	77	52	47	64	72	68	65	73
white	80	76	77	77	73	64	68	76	65	54	73
NLM-WSD dataset	78	88	86	85	81	79	83	85	82	83	85

Conflated CUIs Table 1 shows the accuracy of our conflated CUIs compared to that of the baseline and McInnes et al. McInnes, et al. use the CUIs that are assigned to a term as separate features where conflated CUIs treats them as a single feature. Conflated CUIs obtain an overall accuracy of 85% and McInnes et al. report an overall accuracy of 86% accuracy on the NLM-WSD dataset. The results show there exists no statistically significant difference ($p = 0.5$) in accuracy between the conflated CUIs and the results reported by McInnes, et al. The results also show that conflated CUIs obtains a statistically significant higher accuracy than the baseline ($p = 0.006$).

Ngram CUIs and semantic types Table 1 shows the accuracy of our ngram CUIs and semantic types compared to that of the baseline. The results show that using unigram, bigram and trigram CUIs and semantic types obtain a higher accuracy than the baseline.

Using CUI bigrams and trigrams obtain an accuracy of 81% and 79% respectively. The bigram and trigram accuracy is lower by five and seven percentage points respectively when compared to the accuracy reported by McInnes, et al whose features can be thought of as "CUI unigrams". The results are also not significantly different than the baseline ($p = 0.2$ and $p = 0.4$).

The semantic type bigrams obtain an accuracy of 85% on the NLM-WSD dataset which is comparable to using the CUI unigrams as done by McInnes, et al. There is not a statistically significant difference between these two results ($p = 0.04$).

The semantic type bigrams also obtain a significantly higher accuracy than the semantic type unigrams ($p = 0.0005$). Moving to semantic type trigrams, the results decrease by three percentage points which is also statistically significant $p = 0.00009$).

The results indicated that the accuracy of the classifier does not increase using the ngrams of the biomedical features CUIs but does increase when using the bigrams semantic types.

Global Contextual Features Table 1 shows the accuracy of the global contextual features Journal Descriptors (JDs) and CUIs that obtain a high MMI score for the abstract containing the target word (MMI).

The results show that using the MMI feature obtains an accuracy of 85% on the NLM-WSD dataset, McInnes, et al. report an accuracy of 86% and Stevenson, et al. report an accuracy of 88%. There does not exist a statistically significant difference between the MMI results and those reported by McInnes, et al ($p = .5$) indicating that using the MMI score to select the features does not affect the accuracy of the classifier. There is a significant difference between the MMI results and those reported by Stevenson, et al. ($p = 0.0004$). Although, Stevenson, et al. report an accuracy of 82% when using only the MeSH terms as features into their system, and an accuracy of 88% when combining the MeSH features with linguistic features. We believe that combining MMI and the linguistic features used by Stevenson, et al may increase the accuracy of MMI so that the two classifiers are comparable. The advantage of using MMI over MeSH is that the MMI features is that the MMI features are automatically assigned where as the MeSH feature are manually assigned. MMI features can be used for any piece of text whereas the MeSH feature can only be used for disambiguating words in Medline abstracts.

The results also show that using the Journal Descriptor as a feature obtains an accuracy of 83% on the NLM-WSD dataset. The JD results are interesting even though they are not as good of a feature as CUIs or MeSH terms. JD's are a very high level feature. These are descriptors assigned to the journal that the article containing the ambiguous word. The results show that the subdomain in which the ambiguous word is used is a very good indicator to its concept. For example, $resistance$ in an article in a psychology journal on amnesia would refer to the concept "Psychotherapeutic Resistance" where as $resistance$ in an article in a health informatics journal on opposition to changing lifestyle would refer to "Social Resistance".

6 Unsupervised WSD

There has been very little previous work on unsupervised word sense disambiguation in the biomedical domain that we are aware of although much work has been done in the general English.

SenseRelate (Banerjee and Pedersen, 2003) is an unsupervised word sense disambiguation algorithm that uses the the semantic similarity between the words surrounding an ambiguous word and the possible concepts associated with the ambiguous word. For example, given the sentence "Transport of glutathion conjugates", transport has two possible concepts: i) Biological transport and ii) patient transport. A semantic similarity score is created for each of the concepts by summing the semantic similarity between the concept and each of the concepts associated with the words in the sentence. The concept that has the highest semantic similarity score is assigned to the ambiguous word.

SenseClusters [1] is an unsupervised knowledge-lean word sense disambiguation package The package uses clustering algorithms to group similar instances of target words and label them with the appropriate concept. A vector is created for each concept by taking the centroid of its associated cluster. We refer to these vectors as concept vectors. A vector is created for the target word which we refer to as the target word vector. The cosine is calculated between the target word vector and each of the concept vectors. The concept whose vector is closest to the target word vector is assigned to the target word.

The clustering algorithms in the SenseCluster package include Agglomerative, Graph partitional-based, Partitional biased agglomerative and Direct k-way clustering. The clustering can be done in either vector space where the vectors are clustered directly or similarity space where vectors are clustered by finding the pair-wise similarities among the contexts. The feature options available are first and second-order co-occurrence, unigram and bigram vectors. First-order vectors are highly frequent words, unigrams or bigrams that co-occur in the same window of context as the target word. Second-order vectors are highly frequent words that occur with the words in their respective first-order vector.

In this report, we compare our approach to SenseClusters v0.95 using direct k-way clustering with the I2 clustering criterion function and cluster in vector space. The vectors are created using second-order co-occurrences with a Log Likelihood Ratio greater than 3.84 and the exact and gap cluster stopping parameters (Purandare and Pedersen, 2004; Kulkarni and Pedersen, 2005).

Humphrey et al.(Humphrey et al., 2006) introduce an unsupervised vector approach using Journal Descriptor (JD) Indexing (JDI) which is a ranking algorithm that assigns JDs to journal titles in MEDLINE. The authors apply the JDI algorithm to STs with the assumption that each possible concept has a distinct ST. In this approach, an ST vector is created for each ST by extracting associated words from the UMLS. A target word vector is created using the words surrounding the target word. The JDI algorithm is used to obtain a score for each word-JD and ST-JD pair using the target word and ST vectors. These pairs are used to create a word-ST table using the cosine coefficient between the scores. The cosine scores for the STs of each word surrounding the target word are averaged and the concept associated with the ST that has the highest average is assigned to the target word.

In the following sections, we first discuss the evaluation data used to evaluate the previous approaches discussed above as well as our own. We then discuss our unsupervised approach and our results.

7 Unsupervised Methods

Training Dataset: We use the abstracts from the 2005 Medline Baseline as training data. The data contains 14,792,864 citations from the 2005 Medline repository. The baseline contains 2,043,918 unique tokens and 295,585 unique concepts.

Evaluation Dataset: We use the National Library of Medicine's Word Sense Disambiguation (NLM-WSD) dataset developed by (Weeber et al., 2001) as our test set. This dataset contains 100 instances of 50 ambiguous words from 1998 MEDLINE abstracts. Each instance of a target word was manually disambiguated by 11 human evaluators who assigned the word a CUI or "None" if none of the CUIs described the concept. (Humphrey et al., 2006) evaluate their approach using a subset of 13 out of the 50 words whose majority sense is less than 65% and whose possible concepts do not have the same ST. Instances tagged as "None" were removed from the dataset. We evaluate our approach using these same words and instances.

[1] http://senseclusters.sourceforge.net/

Unsupervised WSD Approach: Our approach has three stages: i) we create a the feature vector for the target word ($instance\ vector$) and each of its possible concepts ($concept\ vectors$) using SenseClusters, ii) we calculate the cosine between the instance vector and each of the concept vectors, and iii) we assign the concept whose concept vector is the closest to the instance vector to the target word.

We explore using the same feature vector parameters as described in the clustering system SenseCluster: i) first-order unigrams, and ii) second-order bigram. To create the the instance vector, we use the words that occur in the same abstract as the target word as features . To create the concept vector, we need to obtain a context that sufficiently represents the meaning of a concept for the task of WSD. The main focus of our unsupervised WSD research has been to answer the question *how to create a vector that can represent the meaning of a concept for unsupervised word sense disambiguation*.

To answer this question, we explore information in the UMLS that can be used to represent the meaning of the concept.

First, we use the concept's CUI definition to represent the meaning of the concept. For example, the term $adjustment$ has three possible concepts in the UMLS version 2008AA: i) C0376209: Individual Adjustment, ii) C0456081: Adjustment Action, and iii) C0683269: Psychological adjustment; each with its corresponding definition. The problem we ran into is that not all CUIs in the UMLS have an associated definition. For example, the term $blood\ pressure$ has three possible concepts: i) C0005823: Blood Pressure, ii) C0005824: Blood Pressure Determination, and iii) C0428878: Arterial pressure. Only the first two though have a corresponding definition.

Due to the fact that not all concepts have a corresponding definition, we explore a backoff model. We use the definition of the concept but if one does not exist we use the definition of its parent to represent its meaning. Here again though, we ran into the same problem; not all parent concepts have an associated definition in the UMLS. We then explore using the semantic type definition to represent the meaning of the concept if its definition does not exist in the UMLS. All semantic types in the UMLS have a corresponding definition.

Second, we use the synonymous terms associated with the concept in the UMLS. For example, the concept C0430400: (Laboratory Culture) has the following synonyms: i) laboratory culture, ii) microbial culture, and iii) sample culture.

Third, we use the terms associated with the siblings of the concept in the UMLS. For example, the concept C0010453 (Anthropological Culture), has the follow terms associated with some of its sibling concepts: i) archeology, ii) family, and iii) social groups.

Lastly, we looked at using the top 50 most frequent words that surround the terms associated with the concept in the 2005 baseline. For example, C0430400: (Laboratory Culture) has the following synonyms: i) laboratory culture, ii) microbial culture, and iii) sample culture.

To summarize, we explore using the following contexts to represent the meaning of a possible concept:

- the concepts definition (CUI)
- the concepts parents definition(s) if the concepts definition does not exist (PAR-DEF)
- the concepts semantic type definition if the concepts definition does not exist (ST-DEF)
- the synonymous terms associated with the concept in the UMLS (SYN)
- the terms associated with the siblings of the concept in the UMLS (SIB)
- the top 50 most frequent words surrounding the terms associated with the concept in the 2005 Medline baseline (TOP 50)

8 Unsupervised Results and Discussion

Table 2 shows the accuracy (%) of our approaches using the following context to represent the meaning of the concept: i) the UMLS definition associated with the possible concept; referred to as CUI, ii) the top 50 most frequent words surrounding the terms associated with the concept; referred to as TOP 50, iii) the synonymous terms associated with the possible concept; referred to as SYN, iv) the definition of the concepts parents in the CUI definition did not exist, v) the definition of the concepts semantic type(s) if the CUI definition did not exist, and iv) the

Table 2: **Accuracy Our Unsupervised Approach on the NLM-WSD Dataset**

target word			1st order		2nd order					
	sense clusters	Humphrey et al.	CUI	TOP 50	CUI	TOP 50	SYN	PAR-DEF	PAR-ST	SIB
adjustment	55	77	54	43	71	68	61	61	68	63
association	0		53	47	64	62	58	58	66	57
blood pressure	54	42	53	47	73	69	60	61	69	59
cold	49		53	44	73	69	58	62	70	60
condition	98	93	53	48	68	68	57	58	71	59
culture	89	100	51	48	73	70	61	62	70	60
degree	97	98	57	61	69	77	78	73	62	63
depression	100	95	55	50	70	77	73	67	59	65
determination	100	100	55	48	68	67	59	59	70	61
discharge	99	93	52	49	74	71	62	63	71	61
energy	99	70	61	65	72	82	84	79	66	67
evaluation	50	60	53	48	69	69	60	59	73	60
extraction	94	98	58	59	75	82	85	77	60	82
failure	86	94	72	85	80	80	80	80	47	80
fat	97	75	53	46	63	62	58	58	66	58
fit	100	100	51	45	65	66	57	60	69	59
fluid	100	60	61	47	74	66	64	64	65	62
frequency	100	90	52	45	66	65	57	60	68	59
ganglion	93	94	52	51	74	72	64	61	72	62
glucose	91	39	53	47	70	70	61	59	73	61
growth	63	70	56	61	74	80	82	76	62	76
immunosuppression	59	75	53	47	64	62	58	58	66	57
implantation	83	94	54	49	66	68	57	59	69	59
inhibition	99	99	55	44	71	68	64	63	68	63
japanese	92	55	94	94	94	94	94	94	78	94
lead	93	39	59	61	76	83	85	78	62	79
man	88		51	54	67	76	72	71	57	62
mole	99	98	53	47	72	70	60	64	70	62
mosaic	54	68	52	48	67	67	59	57	71	58
nutrition	51	35	52	46	63	63	57	57	66	58
pathology	86	75	53	56	66	77	79	74	58	65
pressure	100	12	58	49	72	71	68	67	63	60
radiation	6	79	67	52	74	79	80	73	65	80
reduction	82	100	52	51	71	72	63	60	72	62
repair	76	86	67	58	72	80	83	74	65	79
resistance	100	100	56	44	72	67	68	62	67	61
scale	100	60	52	46	64	64	58	59	67	58
secretion	99	94	57	44	74	68	67	66	67	64
sensitivity	96	83	69	85	95	95	95	95	49	95
sex	80		56	44	70	67	68	62	66	61
single	99	100	59	43	70	68	67	66	67	64
strains	99	98	76	54	87	87	87	87	66	87
support	60	100	52	46	64	64	58	59	67	58
surgery	98	93	53	46	66	65	57	60	68	59
transient	99	99	51	49	70	69	63	58	73	60
transport	99	98	73	62	79	83	84	79	72	84
ultrasound	84	81	53	49	67	69	58	58	70	60
variation	80	73	52	48	67	69	59	57	71	59
weight	55		52	46	64	64	58	59	67	58
white	54	55	52	46	64	64	57	58	67	58
NLM-WSD dataset	83		57	52	71	72	67	66	67	65
Humphrey subset	86	79	57	52	71	72	68	66	67	66

terms associated with the concepts sibling. The table shows all of the representations of meaning using second-order co-occurrence vectors and the results for CUI and TOP 50 using first-order co-occurrence vectors. The $p\text{-}values$ reported are calculated using the one-sided pairwise t-test. In this section, we discuss the results of our approach.

Co-Occurrence Vector Results Table 2 shows the accuracy of representing the meaning of a concept using the CUI definition concept and the TOP 50 most frequent words surrounding the terms associated with the concept using both the first and second-order co-occurrence vectors.

The results show that for both CUI and TOP 50 the second order vectors obtain an accuracy of approximately 20 percentage points higher the first-order vectors. The difference in accuracy between the first-order and second-order vectors is statistically significant for both CUI and TOP 50 ($p = 0.00000009$).

Meaning Representation Results Table 2 shows the accuracy of the different representations of meaning using second-order co-occurrence vectors. The results show that using the TOP 50 and the CUI representations obtain the highest overall accuracy evaluated on the NLM-WSD dataset. TOP 50 obtains an overall accuracy of 72% and CUI obtains an overall accuracy with a precision of 71%. The difference in accuracy between these two results is not statistically significant ($p = .04$).

Using the parent definition to represent the meaning of the concept if the concept definition did not exist in the UMLS (PAR-DEF) shows a significant decrease in accuracy compared to the results of only using the concepts definition (CUI). Similarly, using the semantic type definition to represent the meaning of the concept if the concept definition did not exist also shows a decrease in accuracy over using just the concepts definition CUI. CUI obtains an accuracy of 71% where PAR-DEF obtains an accuracy of 66% while ST-DEF obtains an accuracy of 65%. The addition information reduces the accuracy of the overall results by five and six percentage points respectively. The difference in the results is statistically significant ($p = 0.000009$;$p = 0.003$). We believe that this is the case because the parent and semantic type definitions are too broad to accurately define the meaning of a concept for WSD.

The SYN results obtain an accuracy of 67%; five percentage points lower than using the TOP 50 representation. Analysis of these results show that the synonymous terms associated with the concept are to narrow to represent its meaning. For example, the concept C0456081 (Adjustment Action) has the following synonymous terms: i) adjustment - action, ii) adjustments, iii) adjustment, nos, iv) adjustment - action qualifier value, and v) adjustment - action procedure.

The SIB results also obtain an accuracy of 67%. Analysis of these results show that the terms associated with a concepts siblings are to broad to represent its meaning. For example, the concept C0456081 (Adjustment Action) has approximately 1,350 terms associated with its siblings. Some of those terms are: i) biopsy, ii) cauterisation, iii) cutery, iv) cold therapy, v) desiccation, and vi)state of consciousness.

Comparison Results Table 2 shows the results of our best performing approach that uses the TOP 50 representation with second-order co-occurrence vectors, SenseClusters and Humphrey, et al's vector approach. The difference between the three approaches is how the vectors are created. SenseClusters uses second-order co-occurrence vectors and a clustering algorithm on a set of unannotated training data to create the vectors for each possible concept of the ambiguous word. Humphrey, et al. uses the Journal Descriptor Indexing algorithm and the terms associated with the semantic type of the possible concepts to represent their meaning. Our approach uses second-order co-occurrence vectors and the TOP 50 most frequent words surrounding the terms associated with the concept to represent its meaning.

The results show that SenseClusters obtains the highest accuracy of 86%, Humphrey et al. obtain an accuracy of 79% and our approach obtains an accuracy of 72%.

9 The State of Unsupervised and Supervised WSD

As previously mentioned, the disadvantage to supervised approaches is that they require manually annotated training data for each word that needs

Table 3: **Overall Accuracy of WSD Approaches on different Subsets of the NLM-WSD Dataset**

	Approach	NLM-WSD	Liu Subset	Joshi Subset	Leroy Subset	Humphrey Subset
Supervised	Liu, et al. 04		86			
	Leroy, et al. 05				66	
	Joshi, et al. 05		85	83	77	
	McInnes, et al. 07	86	80	75	82	85
	Stevenson, et al. 08	88	85	83	79	88
Unsupervised	Humphrey, et al. 06				68	79
	SenseClusters	83	82	77	58	86
	Our Unsupervised	72	73	73	62	72

to be disambiguated making them unscalable for large scale systems. Historically though supervised approaches have obtained a higher disambiguation accuracy than unsupervised. Results for unsupervised approaches have been in the 30s (Kilgarriff and Rosenzweig, 2000); meaning that you have a much better chance of picking a sense at random than relying on the algorithm.

Table 3 shows the results of supervised and unsupervised approaches that have been evaluated on different subsets of the NLM-WSD dataset. The results of the supervised and unsupervised approaches on the Leroy subset show that the unsupervised approach described by Humphrey, et al. report an accuracy of 68% while the supervised approach described by Leroy, et al. reports an accuracy of 66%. The supervised system obtains an accuracy two percentage points lower than the supervised system although the difference between the two is not statistically significant ($p = .02$).

The supervised and unsupervised results on the entire NLM-WSD dataset show that the highest performing unsupervised approach, SenseClusters obtains an accuracy of 86% and the best supervised performing system described by Stevenson, et al. reports an accuracy of 88%. The unsupervised system obtains an accuracy only three percentage point lower than the supervised system. The difference in the results is not statistically significant ($p = 0.13482$).

In conclusion, overall the supervised systems still generally obtain a higher disambiguation accuracy than unsupervised, we can see that this is slowly starting to change.

9.1 Future Work

In the future, we plan to explore incorporating semantic similarity measures in our supervised and unsupervised approaches. We are currently in the middle of creating a package called UMLS-Similarity. This package can be used to obtain the semantic similarity between concepts in the UMLS.

There exist a number of *similarity measures* that have successfully used with the ontology WordNet (Fellbaum, 1998) and are implement in the WordNet::Similarity package (Pedersen et al., 2004). Some of these measures were successfully ported to be used with the SnoMed terminology (Pedersen et al., 2006) which is now part of the UMLS. These similarity measures can be grouped into two categories: path length based and information content (IC) based. The path length based measures used are: Wu & Palmer, Leacock & Chodorow, and Path. The IC measures used are: Resnik, Jiang & Conrath, and Lin.

The Path measure is a simple measure that determines the similarity between two concepts by counting the number of nodes between them. This measure was used by (Rada et al., 1989) using the UMLS Metathesaurus.

The (Wu and Palmer, 1994) similarity measure defined in Equation 1 measures the depth of two concepts and the depth of their least common subsumer (LCS). The LCS is the most specific concept two concepts share as an ancestor.

$$sim_{wup} = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \quad (1)$$

The (Leacock and Chodorow, 1998) measure defined in Equation 2 is the negative log of the shortest path between two concepts in a taxonomy divided by twice the total depth of the taxonomy (D).

$$sim_{lch} = -log\frac{minpath(c_1, c_2)}{2 * D} \quad (2)$$

The (Resnik, 1995) measure defined in Equation 3 is based on the information content. It is the negative log of the probability of the concepts which is define as the information content of the LCS of the two concepts.

$$sim_{res} = IC(lcs(c_1, c_2) = -log(P(lcs(c_1, c_2))) \quad (3)$$

The (J. Jiang, 1997) measure defined in Equation 4 is based on the IC of the two concepts as defined by (Resnik, 1995). The measure is modified to include the length of the path between the two concepts.

$$sim_{jcn} = \frac{1}{IC(c_1) + IC(c_2) - 2 * IC(lcs(c_1, c_2))} \quad (4)$$

The (Lin, 1998) measure defined in Equation 5 is also based on (Resnik, 1995) IC but is modified to include the IC of the two concepts.

$$sim_{lin} = \frac{2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (5)$$

We are still considering how this information may be useful in our supervised approach. In our unsupervised approach, we have a number of different ideas.

We plan to use the semantic similarity scores rather than the frequency in the first-order co-occurrence vectors. To create a single vector from the first-order co-occurrence vectors, the vectors are averaged to create a single vector. We plan to explore summing the vectors containing the semantic similarity scores rather than averaging.

We also plan on exploring creating the second-order co-occurrence matrices based on highly similar concepts rather than words from an unannotated corpora. So rather than having an NxN matrix containing the words seen in the corpus, we would use an NxN matrix containing the concepts in the UMLS. And rather than having the cells contain the frequency between the words, they would contain the semantic similarity between the concepts.

Lastly, we plan on exploring using the terms associated with concepts that have a high semantic similarity score with the possible concept to represent the possible concepts meaning.

We believe that by using semantic similarity measures we may catch words or concepts that may be good indicators to the sense of the ambiguous word but do not frequently occur together in the training data. For example, the words $culture$ and $ethnology$. Culture has two possible concepts in the UMLS: Anthropological culture and Laboratory culture. Ethnology is the study of a form of anthropology but it occurs with the term culture only five times in the 2005 Medline baseline which is not a high enough frequency to be considered a relevant feature in our current supervised or unsupervised system. The semantic similarity between the concept Anthropological Culture and the concept of ethnology would be greater than the similarity between Laboratory Culture and the concept of ethnology making a potentially good disambiguation feature.

10 Conclusion

In this report, we discussed two approaches to word sense disambiguation: supervised and unsupervised. We introduce our supervised and unsupervised approaches that we have have explored while participating the National Library of Medicine's Research Participation Program. Historically supervised approaches to word sense disambiguation have obtained a higher accuracy than unsupervised. The results of our supervised and unsupervised work show that this is changing. The unsupervised approaches are obtaining a disambiguation accuracy that are comparable to that of supervised approaches.

Acknowledgments

The author thanks Lan Aronson for allowing me to work with him during my time here. His group François Lang, Jim Mork, Aurélie Névéol and Will Rogers. As well as: Olivier Bodenreider, Allen Browne, May Chey, Guy Divita, Dina Demner-Fushman, Kin Wah Fung, Susanne Humphrey, Dwayne McCully, Tom Rindflesch and Suresh Srinivasan.

Our experiments were conducted using

CuiTools v0.19, which is freely available from http://cuitools.sourceforge.net.

References

A Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 17–21.

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810.

C Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press USA.

Susanne M. Humphrey, Willie J. Rogers, Halil Kilicoglu, Dina Demner-Fushman, and Thomas C. Rindflesch. 2006. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *J. Am. Soc. Inf. Sci. Technol.*, 57(1):96–113.

S.M. Humphrey. 1999. Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society for Information Science*, 50(8):661–674.

D. Conrath J. Jiang. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics, Taiwan*, pages pp. 19–33.

M. Joshi, T. Pedersen, and R. Maclin. 2005. A comparative study of support vectors machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of Second Indian International Conference on Artificial Intelligence*, pages 3449–3468.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34(1):15–48.

A. Kulkarni and T. Pedersen. 2005. SenseClusters: unsupervised clustering and labeling of similar contexts. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 105–108, June.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. pages 265–283.

G. Leroy and T.C. Rindflesch. 2005. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *International Journal of Medical Informatics*, 74(7-8):573–585.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

H. Liu, V. Teller, and C. Friedman. 2004. A multi-aspect comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association*, 11(4):320–331.

B. McInnes, T. Pedersen, and J. Carlis. 2007. Using umls concept unique identifiers (cuis) for word sense disambiguation in the biomedical domain. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 533–37, Chicago, IL, Nov.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Demonstration Papers*, pages 38–41, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

T. Pedersen, S.V. Pakhomov, S. Patwardhan, and C. Chute. 2006. Measures of semantic similarity and relatedness in the biomedical domain. *Biomedical Informatics, Elsevier*.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48.

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

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453.

M. Stevenson, Y. Guo, R. Gaizaskas, and D. Martinez. 2008. Knowledge sources for word sense disambiguation of biomedical text. In *Proceedings of the BioNLP workshop*, Columbus, Ohio, July. Association for Computational Linguistics.

M. Weeber, JG. Mork, and AR. Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 746–750.

I.H. Witten and E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection.