

Accurate and Scalable Word Sense Disambiguation in the Biomedical Domain

A THESIS PROPOSAL

BY

Bridget T. McInnes

FOR A DOCTORAL PRELIMINARY ORAL EXAMINATION

FOR THE DEGREE OF

PHILOSOPHY OF SCIENCE

AND

FOR THE ENJOYMENT OF

ALL INVOLVED

January 2007

Contents

1	Introduction	1
2	Data and Resources	8
2.1	General English Test Data sets	8
2.1.1	“interest”, “line”, “hard”, and “serve” Data Sets	8
2.1.2	SemCor Data Set	8
2.1.3	Senseval Data Sets	9
2.2	Biomedical Test Data set	9
2.2.1	NLM-WSD Data Set	9
2.3	External Knowledge Sources	10
2.3.1	Unified Medical Language System (UMLS)	10
2.3.2	WordNet	11
3	Related Work	12
3.1	Supervised WSD Systems	12
3.1.1	Supervised WSD Algorithms	12
3.1.2	Supervised WSD Features	14
3.1.3	Supervised WSD Results	18
3.2	Unsupervised WSD Systems	22
3.2.1	Clustering	22
3.2.2	Parallel Text-based	24
3.2.3	Unsupervised WSD Results (Accuracy)	25

3.3	Knowledge-based WSD Systems	29
3.3.1	Similarity and Relatedness-based	29
3.3.2	Contextual Knowledge-based	34
3.3.3	Frequency-based	35
3.3.4	Vector-based	38
3.3.5	Knowledge-based WSD Results	40
3.4	Bootstrapping WSD Systems	43
3.5	Overview of WSD Systems	44
4	Preliminary Work	45
4.1	UMLS Concepts	45
4.2	Journal Descriptors	46
4.3	Results and Discussion	46
4.3.1	Surrounding UMLS Concepts	46
4.3.2	Journal Descriptors	50
4.3.3	Algorithm Results	51
4.4	Preliminary Conclusion	51
5	Proposed Work	54
5.1	MetaMap	56
5.2	WSD-enhanced MetaMap System	57
5.3	Evaluation of Proposed Knowledge-based WSD System	59
5.4	Possible Ways Forward	59
5.4.1	Knowledge Source Information	60

5.4.2	Knowledge-based WSD Algorithm	61
5.5	Proposed Contributions	61

List of Figures

1	Simplified WSD System Diagram	2
2	Supervised WSD System Diagram	3
3	Unsupervised WSD System Diagram	4
4	Knowledge-based WSD System Diagram	5
5	Bootstrapping WSD System Diagram	7
6	Qualitative Analysis of WSD Systems	44
7	Scalability versus Accuracy of WSD systems	54
8	Proposed Knowledge-based WSD System	55
9	Current MetaMap System	56
10	WSD-enhanced MetaMap System	58

List of Tables

1	Supervised WSD Algorithms	13
2	Supervised WSD Features	15
3	Supervised WSD Results	21
4	Unsupervised WSD Features	23
5	Unsupervised WSD Results	26
6	Semantic and Relatedness Measures used in Knowledge-based WSD Systems	33
7	Knowledge-based WSD Results	41
8	Qualitative Analysis of WSD Systems	44
9	UMLS Concept Results using Naive Bayes	48
10	UMLS Concept Results using SVMs	49
11	Journal Descriptor Results using Naive Bayes	50
12	Supervised WSD Algorithm Results using Naive Bayes and SVM	52
13	Overall Feature Results	53
14	Overall Algorithm Results	53

1 Introduction

Word Sense Disambiguation (WSD) is the task of identifying the appropriate sense of a word that has multiple senses. This word is referred to as the target word. WSD has been deemed an important problem. For example, (Rigau et al., 2002) state “WSD is one of the most important open problems in Natural Language Processing (NLP).”, (Liu, Teller, and Friedman, 2004) state “Resolving sense ambiguity is one of the most important problems in Natural Language Processing.”, and (Yarowsky, 1992) states “Word sense disambiguation is a long-standing problem in computational linguistics with important implications for a variety of practical applications”. Research in this area started in the 1940s, yet still, “no large-scale broad-coverage accurate WSD system has been built up to date with current state-of-the-art accuracy in the range of 60-70%” (Rigau et al., 2002)¹.

WSD applies to a number of NLP tasks such as “text-to-speech, machine translation, information retrieval” (Gale, Church, and Yarowsky, 1992c) and concept mapping (Aronson, 2001).

Text-to-speech is the task of producing the speech equivalent of written text. An example of such a system is an automatic announcement system for the weather, airport arrivals/departures, or movie showings. The appropriate sense of a word is needed to pronounce some words properly. For example the word “bass”, which is pronounced as [beys] to mean a low pitched singing voice or [bæs] to mean the fish.

Machine translation is the task of translating a text from one language into another, such as English to French. The appropriate sense of a word is needed to translate it properly. For example, the French word ‘grille’ can be translated to ‘railings’, ‘gate’, ‘bar’, ‘scale’ or ‘schedule’ depending on its context (Ide and Véronis, 1998).

Information retrieval is the task of indexing, searching, and recalling data. (Ide and Véronis, 1998) states “When searching for keywords, it is desirable to eliminate occurrences in documents where the word or words are used in an inappropriate sense”. In order for this to occur, documents need to be properly indexed based on the sense of the words in the documents rather than the word itself. For example, the word “fluke” can mean a flatfish, the end part of an anchor, the fins on a whale’s tail or a stroke of luck.

Concept mapping is the task of automatically linking documents to concepts (senses) in an ontology. This

¹the accuracy comes from the Senseval-2 WSD task (see Section 2)

is done by linking content words in documents to their appropriate concept in the ontology. In order to do this accurately, the appropriate concept needs to be identified. This is a very similar problem to that of information retrieval. MetaMap (Aronson, 2001) is a concept mapping system that maps Medline articles to concepts in the Unified Medical Language System (see Section 2). Medline is an online database that contains 11 million references to journal references to biomedical articles. Metamap determines the appropriate concept through a pattern matching algorithm using lexical variation of the input words. (Aronson et al., 2000) and (Aronson, 2001) note that a WSD component would greatly improve the accuracy of the MetaMap system’s mappings. (This will be discussed in further detail in Section 5.)

The remainder of this section consists of a high level overview of WSD systems. Section 2 presents data and resources used by WSD systems. Section 3 presents in detail WSD related work. Section 4 presents our preliminary work and Section 5 presents our proposed work.

Researchers differ in their definitions of what constitutes a word. In the NLM-WSD data set, the target word may contain more than one word, such as “blood pressure”. In all the other data sets discussed in this proposal, a target word consists of only a single word, such as “cardiology”. When we discuss “surrounding words”, we are referring to single words that surround the target word regardless of the data set. We also do not include abbreviations such as “AARP” or acronyms such as “e.g.” in our definition of words. We are making the assumption that these have already been identified and expanded.

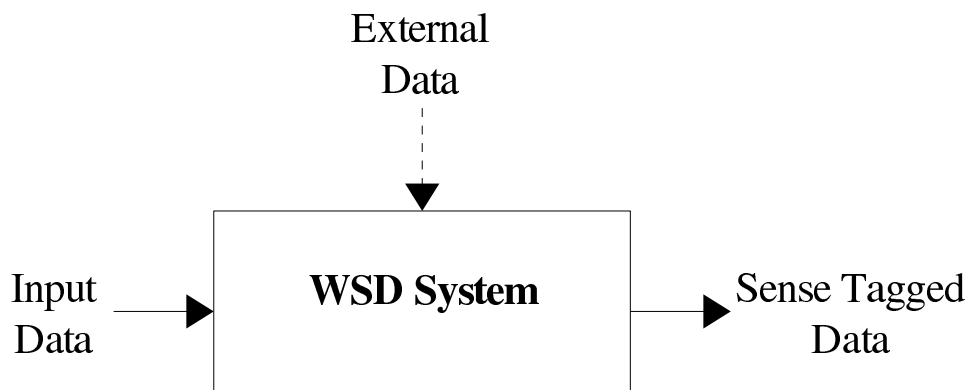


Figure 1: Simplified WSD System Diagram

Figure 1 shows a simplified diagram of a WSD system. It takes a data set as input (input data) and possibly external data and returns the data set with all words in the data set annotated with their respective senses (sense tagged data). Continuous arrows represent system requirements while dashed arrows are sometimes

required. An example of input data would be a sentences from Medline articles or abstracts such as “... expression to lower blood pressure ...”. The WSD system identifies the sense of the words in the input data possibly using information from external data such as a corpus (e.g. British National Corpus, GigaWord Corpus) or knowledge source (e.g. WordNet, Macquarie’s Thesaurus, Unified Medical Language System). The WSD system assigns a sense to each word in the input data returning sense tagged data such as “... <calcium/C0006675> <to> <lower/C1611820> <blood pressure/C1272641> ...” where C0006675, C1611820, and C1272641 are sense identifiers that come from a sense inventory inside the system. A sense inventory contains a list of words and their possible senses. “None” is typically included as a possible sense even though it is not an actual sense. It is used to indicate that none of the senses apply to the target word.

We organize WSD systems into four categories: supervised, unsupervised, knowledge-based, and bootstrapping. Next, we give an overview of each of the systems.

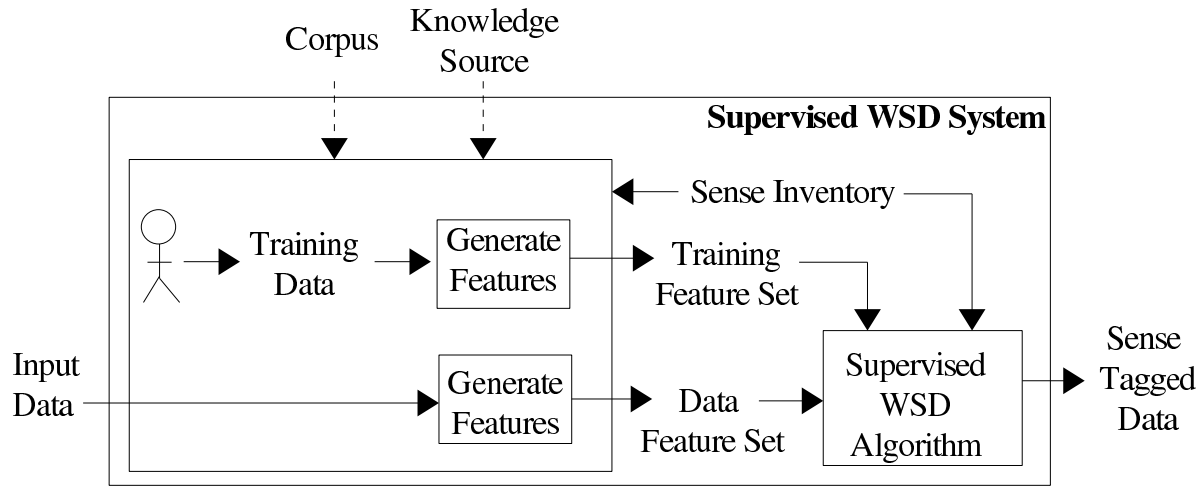


Figure 2: Supervised WSD System Diagram

Figure 2 shows a supervised WSD system which uses manually sense-tagged training data and possibly an external corpus or knowledge source. A module in this system generates a feature set of both the training data and the ambiguous words (target words) in the input data. External corpora and/or external knowledge sources are sometimes used to obtain the features. A feature set contains a feature vector for each instance of the target word. An instance is a single example in a text of the target word. There may exist multiple instances of a target word. Target word also refers to the specific word that we are trying to disambiguate in an instance. For example, we are trying to disambiguate the word “lower” in the instance: “He wants lower

blood pressure.”. The sentence is an instance of the target word “lower”, and “lower” is the target word in the instance. A feature vector represents an instance as an n-dimensional vector of numerical features. Examples of a feature include: the part-of-speech (POS) of the target word, the words surrounding the target word, and the number of times the target word occurs with the words to its left and right. Supervised WSD algorithms require training feature vectors in order to group them in an abstract space based on their sense. This is called the feature space. The idea is that feature vectors whose target words have the same sense will be situated close together inside the feature space. An input data feature vector of a specific target word will be assigned the sense of the training feature vectors it is closest to. The algorithm assigns a sense to each instance of a target word in the input data feature set from a given sense inventory and outputs the sense tagged data.

The problem with a supervised WSD system is that there must be sufficient training data for each word that needs to be disambiguated in the input data. The training data for a supervised WSD system is manually created by humans making the system not scalable because it is intractable for large scale problems such as Metamap where every content word in the Medline article needs to be mapped to a UMLS concept (sense). (Joshi, Pedersen, and Maclin, 2005) state that a supervised WSD system is “not scalable due to the time and effort involved in manually creating sufficient amounts of training data”.

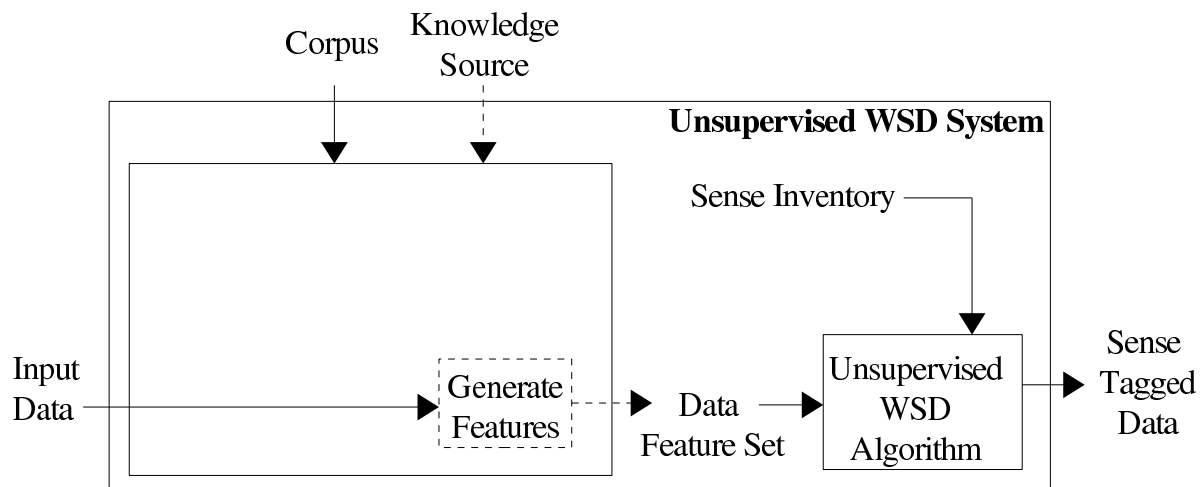


Figure 3: Unsupervised WSD System Diagram

Figure 3 shows an unsupervised WSD system which does not require training data, thereby eliminating the problem of needing manually tagged training data for each ambiguous word in the input data. An unsuper-

vised WSD system is “data-driven”, relying on the “distributional characteristics of unannotated corpora and traditional equivalences in word-aligned parallel text” (Agirre and Edmonds, 2006). It is “based on the similarity exhibited among the contexts in which words occur in unannotated corpora” (Agirre and Edmonds, 2006). We subdivide unsupervised WSD systems into knowledge-lean and knowledge-dependent depending on the external data used by the system. A knowledge-lean unsupervised WSD system’s external data only comes from a corpus while knowledge-enhanced system’s are augmented by an external knowledge source.

The problem with an unsupervised WSD system is that a state-of-the-art unsupervised WSD system does not perform as well as a state-of-the-art supervised WSD system. In an analysis of the Senseval-3 English all words WSD task (see Section 2), (Snyder and Palmer, 2004) note that “As with the SENSEVAL-2 English all-words task, the supervised systems fared much better than unsupervised systems. In fact, all the seven systems reported as supervised scored higher than any of the nine systems reported as unsupervised in both precision and recall (using either the two scoring criteria)”.

Figure 4 shows a knowledge-based WSD system which is knowledge-driven. It requires the use of an external knowledge source such as dictionaries, thesauri and ontologies. Such a system does not rely on supervised or unsupervised algorithms but “by using information from an explicit lexicon [corpus] or knowledge base [source]” (Molina et al., 2002). The system does not require a module to generate features sets. The algorithm in the system requires information from a knowledge source and possibly augmented by an external corpus.

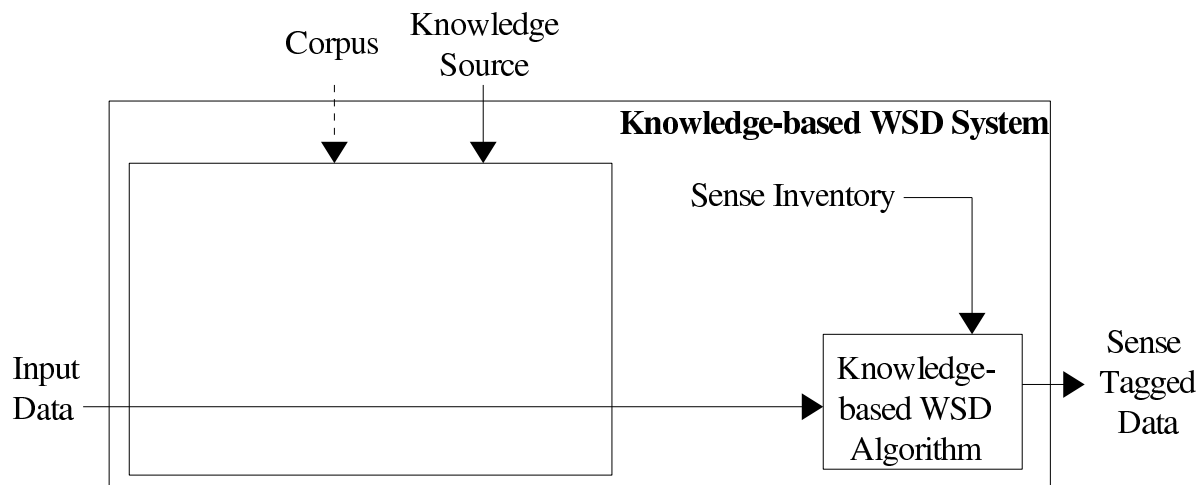


Figure 4: Knowledge-based WSD System Diagram

The problem with a knowledge-based WSD system is the same as that of a unsupervised WSD system, a supervised WSD system perform better. In an analysis of the Senseval-2 English lexical WSD task, (Molina et al., 2002) note that a supervised WSD system “achieve(s) better results than knowledge-based ones”.

Figure 5 shows a bootstrapping WSD system which compensates for the decrease in performance in an unsupervised or knowledge-based WSD systems and alleviates somewhat the problem of a supervised WSD system’s need for training data. A bootstrapping WSD system combines an unsupervised or knowledge-based WSD system and a supervised WSD system into one. The sense-tagged output data of an unsupervised or knowledge-based WSD system feeds in as input data to a supervised WSD system. In some systems the output of a supervised WSD system is iteratively fed back into the training data until a convergence factor is met. Other systems incorporate a small set of manually annotated examples, called seeds, as their external data source. A bootstrapping WSD system performance is comparable to that of a supervised WSD system (Yarowsky, 1992). (Kilgarriff and Rosenzweig, 2000) states that “Where there is training data available, systems that use it perform substantially better than those that do not”.

The problem with a bootstrapping WSD system is that the training data still must be created for each word that is to be disambiguated in the input data. Automatically (or semi- automatically with a small seed set) obtaining the training data is cheaper than manually creating all of it but it is computationally expensive for large scale problems and it is not always

The general goal of the proposed research is to create a scalable WSD system that achieves accuracy that is comparable to a state-of-the-art supervised WSD system.

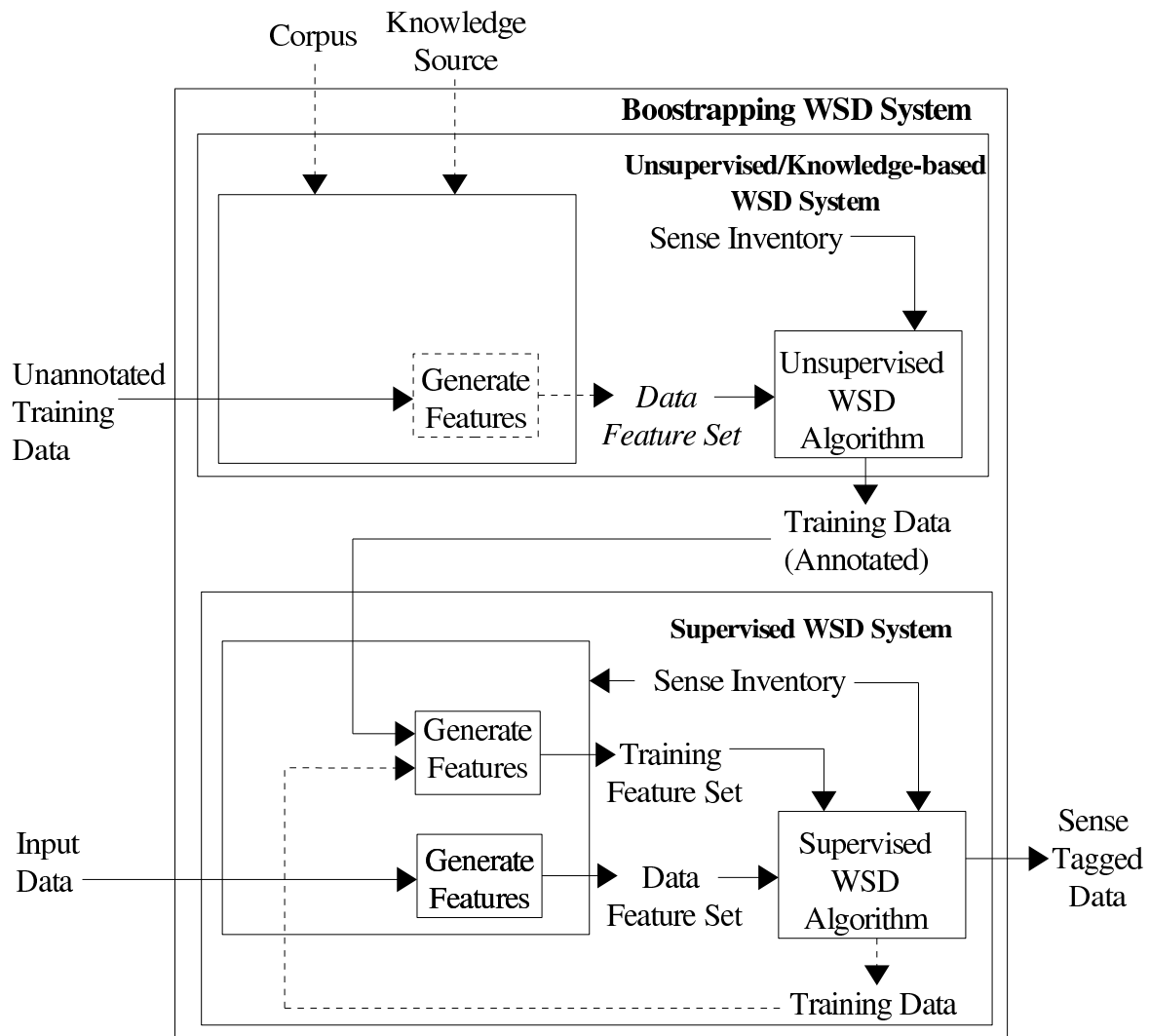


Figure 5: Bootstrapping WSD System Diagram

2 Data and Resources

In this section, we discuss the test data sets in the general English and biomedical domains. We then discuss two external knowledge sources, Unified Medical Language System (UMLS) and WordNet, that have been used by WSD systems.

2.1 General English Test Data sets

2.1.1 “interest”, “line”, “hard”, and “serve” Data Sets

The “interest” data set (Bruce and Wiebe, 1994) contains 2,368 instances of the noun “interest” from a subset of the Penn Treebank Wall Street Journal Corpus (ACL/DCI version). Each instance was manually annotated with one of six senses from the Longman Dictionary of Contemporary English (LDOCE).

The “line” data set (Leacock, Miller, and Chodorow, 1998) contains 4,149 instances of the noun “line” from the 1987-1989 Wall Street Journal Corpus and the American Printing House of the Blind. Each instance was manually annotated with one of the six possible senses from WordNet.

The “hard” data set (Leacock, Miller, and Chodorow, 1998) contains 4,337 instances of the adjective “hard” from the San Jose Mercury New Corpus. Each instance was manually annotated with one of three possible senses from WordNet.

The “serve” data set (Leacock, Miller, and Chodorow, 1998) contains 5,131 instances of the verb “serve” from the 1987-1989 WSJ corpus and the American Printing House for the Blind. Each instance is manually annotated with one of four possible senses from WordNet.

2.1.2 SemCor Data Set

SemCor contains 250,000 words from the Brown Corpus and the novel “The Red Badge of Courage”. The content words were manually tagged using WordNet as the sense inventory. 83 target words have more than 100 sense-tagged instances in the training data.

2.1.3 Senseval Data Sets

Senseval is an international organization whose goal is to promote research in WSD. The organization runs evaluation exercises to test WSD systems. There are three evaluations Senseval-1 which took place in 1998, Senseval-2 which took place in 2001 and Senseval-3 which took place in 2004.

The Senseval-1 data set contains 35 target words with 13,845 training instances and 7,446 test instances that were manually tagged using the sense inventory HECTOR.

The Senseval-2 data set contains Chinese lexical sample, Danish lexical sample, Dutch all words, Czech all words, Basque lexical sample, Estonian all words, Italian lexical sample, Korean lexical sample, Spanish lexical sample, Swedish lexical sample, Japanese lexical sample, Japanese translation, English all words, and English lexical sample. The systems reported in this paper use the English all words (Senseval-2AW) or English lexical sample (Senseval-2LS). The *English lexical sample* contains 73 target words with 8,611 training instances and 4,328 test instances from BNC-2, the Penn Treebank that were manually tagged using the sense inventory WordNet1.7. *English all words* contains a corpus of 2,456 words from the Penn Treebank where all content words in the corpus are manually tagged using the sense inventory WordNet1.7.

Senseval-3 data set contains Italian all words, Basque lexical sample, Catalan lexical sample, Chinese lexical sample, Romanian lexical sample, Spanish lexical sample, multilingual lexical sample, WSD of WordNet glosses, English lexical sample and English all words. The systems reported in this paper used the *English all words* (Senseval-3AW) and *WSD of WordNet glosses* (Senseval-3WN). The Senseval-3AW contains 2,081 words from the Penn Treebank where all the words in the corpus were manually tagged. The Senseval-3WN contains 15,717 words from WordNet glosses (definitions) where each of these words were manually tagged. The sense inventory used for these data sets was WordNet1.7 for nouns and WordSmith for verbs.

2.2 Biomedical Test Data set

2.2.1 NLM-WSD Data Set

National Library of Medicine's Word Sense Disambiguation (NLM-WSD) data set contains 50 highly frequent ambiguous Unified Medical Language System (UMLS) concepts from the 1998 MEDLINE abstracts. Each target word in the NLM-WSD data set contains 100 ambiguous instances randomly selected from the

1998 abstracts totaling to 5,000 instances. The instances were manually disambiguated by 11 evaluators who assigned the target word to a UMLS Concept or assigned the sense as “None” if none of the UMLS concepts described the sense. There are 15 out of the 50 terms whose majority sense is less than 65%. Those terms are used by (Leroy and Rindflesch, 2004) and subsequently our preliminary work.

2.3 External Knowledge Sources

2.3.1 Unified Medical Language System (UMLS)

The UMLS is a knowledge representation framework designed to support broad scope biomedical research queries. It includes over 100 controlled medical terminologies and classification systems encoded with different semantic and syntactic structures. The three major sources of UMLS are the Metathesaurus, Semantic Network and SPECIALIST Lexicon.

The Metathesaurus is a multi-lingual vocabulary database. It contains information about biomedical and health-related concepts, relationships among the concepts, and synonymous terms that are associated with the concepts. The Metathesaurus organizes knowledge based on concepts. A concept is defined as the “meaning” of a term and is expressed by having specific attributes that define it. A concept contains a concept definition, related concepts, relations with other concepts and semantic types defined from the Semantic Network.

The Semantic Network (SN) contains information about a Metathesaurus concept’s semantic type and its relationship with other semantic types. A semantic type is a cluster of words that are meaningfully related in some way. A concept could have more than one semantic type. There are currently 135 semantic types. Examples of semantic types include: organism, anatomical structures, biologic function, and chemicals. The semantic types are connected by 54 semantic relations. Examples of semantic relations include: is-a, part-of, ingredient-of.

The SPECIALIST Lexicon contains English biomedical terms and English terms that are used in the biomedical and health-related domain as well as NLP tools such as the SPECIALIST minimal commitment parser, and lexical variation generator (LVG). A term may consist of more than one word. There exists a lexical entry for each spelling or spelling variation. An entry may have more than one UMLS Concept.

2.3.2 WordNet

WordNet (Fellbaum, 1998) is an ontology that can also be used as a machine readable dictionary (MRD) or thesaurus (MRT). WordNet contains approximately 150,000 words that are grouped together by their concepts called synsets. There are approximately 115,000 synsets. A word-synset pair is defined by its part-of-speech (POS), definition, and a set of example sentences. Synsets are linked together through five semantic relations: *hypernym*, *hyponym*, *meronym*, *holonym* and *synonym*. The hypernym of words w_1 and w_2 is when the meaning of w_1 encompasses the meaning of w_2 . For example, a truck (w_1) is a kind of vehicle(w_2), therefore a truck is a hypernym of a vehicle. The hyponym of words w_1 and w_2 is the hypernym relationship backwards. For example, a truck (w_2) is a kind of vehicle (w_1), therefore a vehicle is a hyponym of a truck. The meronym of words w_1 and w_2 is when w_1 is part of or a member of w_2 . For example, a wheel (w_1) is part of a truck (w_2), therefore a wheel is a meronym of a truck. The holonym of words w_1 and w_2 is when w_1 has an w_2 as a component. For example, a truck (w_1) has a wheel (w_2), therefore a truck is a holonym of a wheel. The synonym of words w_1 and w_2 is when w_1 and w_2 have the same meaning. For example, a *baby* (w_1) is an *infant* (w_2), therefore a baby is an infant.

3 Related Work

WSD systems fall into the four categories described in the previous section: supervised, unsupervised, knowledge-based and bootstrapping. In this section, we conduct a comparative analysis of the work related to our proposed research in Section 5.

3.1 Supervised WSD Systems

There are two components to a supervised WSD system (see Figure 2, Section 1): i) the module that generates feature sets for the training and input data and ii) the algorithm which determines the sense of ambiguous words in the input data feature set using the training data feature set. We first discuss the algorithms researchers use to disambiguate the input data, and second, the features used in the feature set, and lastly, the results of relevant related work.

3.1.1 Supervised WSD Algorithms

Table 1 shows the algorithms used by researchers in supervised WSD. There are ten papers and seventeen algorithms discussed in this section. The papers each compare on average three algorithms on various test sets. The order of the papers in the table is in the order they are discussed below.

(Mooney, 1996) compares Naive Bayes, Neural Networks, Decision Trees, Decision Lists, K-nearest neighbor, logic-based DNF and CNF on the “line” data set. (Leacock, Towell, and Voorhees, 1993) compare Naive Bayes, Neural Network and Content Vector on the same data set. Both found that the Naive Bayes and Neural Networks performance was comparable and performed better than the other algorithms.

(Yarowsky and Florian, 2003) compare Naive Bayes, Decision Lists, Cosine, Transformation based and (Gale, Church, and Yarowsky, 1992b) Bayes Ratio on the Senseval-2LS data set. They found that although the Naive Bayes and Bayes Ratio reports the best overall performance although they point out that no single algorithm always performs the best on all the words to be disambiguated.

(Liu, Teller, and Friedman, 2004) compare the Naive Bayes, a modified Decision List algorithm and their mixed supervised WSD system which contains a mix of a Naive Bayes and an exemplar-based algorithms.

Table 1: Supervised WSD Algorithms

	(Mooney, 1996)	(Leacock, Towell, and Voorhees, 1993)	(Yarowsky and Florian, 2003)	(Liu, Teller, and Friedman, 2004)	(Ng and Lee, 1996)	(Ng, 1997)	(Lee, Ng, and Chia, 2004)	(Joshi, Pedersen, and Maclin, 2005)	(Leroy and Rindfleisch, 2004)	(Lee and Ng, 2002)
Naive Bayes	x	x	x	x		x		x	x	x
SVM							x	x		x
Ada Boost								x		x
Decision Tree	x							x		x
Decision List	x		x	x				x		
Instance-based										
Exemplar-based					x	x				
Exemplar/Naive Bayes “mixed”				x						
Content Vector		x								
Neural Network	x	x								
Ensemble method										
Cosine			x							
Transformation-based			x							
Bayes Ratio			x							
K-Nearest Neighbor	x									
logic-based DNF	x									
logic-based CNF	x									
Data Set										
“line” data set	x	x		x						
“hard” data set				x						
“serve” data set				x						
“interest” data set				x						
Senseval-1										x
Senseval-2LS			x							x
BNC/WSJ6			x		x	x				
NLM-WSD				x				x	x	

They tested the systems in the general English domain using the “line”, “hard”, “serve” and “interest” data and the biomedical domain using a subset of the NLM-WSD data set. They found Naive Bayes performed best in the general English domain and their “mixed” supervised WSD system performed the best in the biomedical.

(Ng, 1997) compare Naive Bayes and an improved version of the exemplar-based learning algorithm, LEXAS (Ng and Lee, 1996) on a subsection of the British National Corpus (BNC) and a subsection of the Penn Treebank Wall Street Journal (WSJ6) corpus. They found no difference in the algorithms performance.

(Joshi, Pedersen, and Maclin, 2005) and (Lee and Ng, 2002) compare Naive Bayes, Support Vector Machines (SVM), AdaBoost and Decision Trees. (Joshi, Pedersen, and Maclin, 2005) also compare Decision Lists. (Joshi, Pedersen, and Maclin, 2005) test their algorithm using a subset of the NLM-WSD data set while (Lee and Ng, 2002) use the Senseval-1 and Senseval-2LS data sets. They both report that SVM obtained the best overall accuracy. (Lee, Ng, and Chia, 2004) evaluate an SVM on the Senseval-3LS data set.

In summary, SVM’s and Naive Bayes were shown to outperform all of other supervised WSD algorithms overall but (Leacock, Towell, and Voorhees, 1993) and (Joshi, Pedersen, and Maclin, 2005) note that no one classifier performed best for all words.

3.1.2 Supervised WSD Features

Table 2 shows the features used by researchers in supervised WSD. There are ten features that we group into three categories: lexical, semantic and syntactic. Lexical features consist of features that can be obtained by analyzing the target word and its surrounding words in the sentence. The features relevant to our discussion include bag-of-words, unigrams, bigrams, and collocations. Syntactic features consist of features that can be obtained by analyzing the syntactic structure of the words and the sentence such as the morphology and part-of-speech of the target and surrounding words, head words in the sentence and syntactic relations between the target word and surrounding words. Semantic features consist of features that require additional knowledge of a word including it’s definition and sense. Semantic features include the semantic type and semantic relations of the target word and possibly the words surrounding the target word.

Table 2: Supervised WSD Features

		(Leacock, Towell, and Voorhees, 1993)	(Mooney, 1996)	(Pedersen, 2000)	(Ng and Lee, 1996)	(Yarowsky and Flo- rian, 2003)	(Lee and Ng, 2002)	(Mohammad and Ped- ersen, 2004)	(Joshi, Peder- sen, and Maclin, 2005)	(Leroy and Rind- flesch, 2004)	(Lee, Ng, and Chia, 2004)
Lexical Features	bag-of-words unigrams bigrams collocations	x	x	x	x	x x x x	x x	x x	x x		x x
Syntactic Features	morphology POS head word syntactic relations				x x x	x x	x x			x x	x x
Semantic Features	semantic types semantic relations									x x	

Researchers have incorporated four lexical features in supervised WSD systems: bag-of-words, unigrams, bigrams, and collocations. A bag-of-words are the words surrounding the target word without respect to order. A unigram is a content word surrounding the target word that occurs a specified number of times in a specified window. The difference between a bag-of-words and unigrams is that the collection of unigrams take the frequency of the words into consideration at some point in the process. A bigram is an ordered pair of content words that surround the target word occurring a specified number of times in a specified window. A window is the number of words on either side of the target word. For example, a window size of three would be one word to the right and one word to the left of the target word. A window could also be the entire sentence that contains the target word. A collocation is a unit of words that represent a single concept for example “White House”.

(Leacock, Towell, and Voorhees, 1993), (Mooney, 1996) and (Pedersen, 2000) use the bag-of-words feature. (Ng and Lee, 1996) compare bag-of-words and collocations while (Lee and Ng, 2002) compare collocations and unigrams with their systems. (Mohammad and Pedersen, 2004) and (Joshi, Pedersen, and Maclin, 2005) compare unigram and bigram from the words surrounding the target word at the sentence level and since the data they use is biomedical abstracts, at the abstract level. (Lee, Ng, and Chia, 2004) use both unigrams and collocations in their supervised WSD system.

Researchers have incorporated four syntactic features in supervised WSD systems: morphology, part-of-speech information (POS), head word information and syntactic relations. The morphology of the target word is the “analysis of each word into its root and affix” (McRoy, 1992), the POS information is the POS of the target word and/or surrounding words, the head word information is whether the target word and/or specified surrounding words are the head words of their respective phrase, and syntactic relations is the relationship between the POS of the target word and the POS of its surrounding words.

(Ng and Lee, 1996) use the morphology as a feature in their system in conjunction with POS and compare it with the verb-object syntactic relation. A verb-object syntactic relation exists if the target word is the head word of a noun phrase and the word immediately preceding the phrase is a verb. (Lee and Ng, 2002) and (Lee, Ng, and Chia, 2004) use the noun syntactic relation between the POS of the target word and the surrounding words. The noun syntactic relation depends on the POS of the word. If it is a noun, there are four features, its parent headword, the POS and voice of the headword and its orientation to the headword. Orientation is the whether the headword is to the left or the right of the target word. (Lee, Ng, and Chia,

2004) also uses the POS of the surrounding words. (Leroy and Rindflesch, 2004) use the POS of the target word and whether the target word is a head word. (Mohammad and Pedersen, 2004) use variations of head word and POS of the target word and surrounding words in their feature sets. The feature sets that returned the best accuracy are i) POS of the specified surrounding words, ii) the head word of the phrase containing the target word, iii) the head word of the target word's parent phrase, iv) the POS of the head word of the phrase containing the target word and v) the POS of the head word of the target word's parent phrase.

Researchers have incorporated two semantic features in supervised WSD systems: semantic type and semantic relations. A semantic type is a broad subject categorization assigned to a sense (UMLS concept) such as the semantic type of the word "pacemaker" is "Medical Device". A sense may have more than one semantic type. A semantic relation is a relationship between semantic types. For example, the semantic relation between the the semantic types "Manufactured Device" and "Medical Device" is "is-a".

(Leroy and Rindflesch, 2004) introduce a supervised WSD system that disambiguates word in MEDLINE abstracts by mapping them to their appropriate sense (concept) in the Unified Medical Language System (UMLS). (Leroy and Rindflesch, 2004) use the following semantic features: i) the semantic types of the words surrounding words the target word at the phrase level and sentence level, and ii) the UMLS Semantic Network relation between the semantic type of the target word and the semantic type of the surrounding words. They authors split the relations into two sets, *core* relations and *non-core* relations due to their hierarchical nature. The *core* relations are hierarchical, for example: *is-a*, *conceptual-part-of*, *consists-of*, *contains*, and *process-of*. All other relations are identified as *non-core* relations. (Leroy and Rindflesch, 2004) use these relations in two different ways. First, they test to see if additional semantic knowledge about the surrounding context of the word will increase the performance. They do this by using the *core* and *non-core* relations between the semantic types of the surrounding words. Second, they use the *core* and *non-core* relations between the semantic type of the target word and the semantic types of its surrounding words.

In summary, feature selection is a difficult problem. (Yarowsky and Florian, 2003) note that there has not been found a set of features that best disambiguates all ambiguous words.

3.1.3 Supervised WSD Results

Table 3 shows the the accuracy of relevant related work and the data sets on which they were evaluated. (Leacock, Towell, and Voorhees, 1993), (Mooney, 1996) and (Pedersen, 2000) evaluate their system on the “line” data set. (Ng and Lee, 1996) evaluate their system on the “interest” data set. (Lee and Ng, 2002) evaluate their system on the Senseval-1 and Senseval-2LS data sets. (Mohammad and Pedersen, 2004) evaluate their systems on the six data sets Senseval-1 and Senseval-2LS, “line”, “hard”, “serve” and “interest”, allowing them to be compared to all of the above. (Leroy and Rindflesch, 2004), (Joshi, Pedersen, and Maclin, 2005) and (Liu, Teller, and Friedman, 2004) evaluate their systems on a subset of the NLM-WSD data set. (Liu, Teller, and Friedman, 2004) also evaluates their system on the “line”, “hard”, “serve”, and “interest” data sets.

(Leacock, Towell, and Voorhees, 1993), (Mooney, 1996) and (Pedersen, 2000) use the bag-of-words feature and achieve an accuracy of 76%, 72%, and 88% respectively. (Mohammad and Pedersen, 2004) report a higher accuracy for “line” data unigram and bigram results than the bag-of-words results reported by (Leacock, Towell, and Voorhees, 1993) and (Mooney, 1996) but not for results reported by (Pedersen, 2000).

(Ng and Lee, 1996) evaluate bag-of-words and collocations on the “interest” data set finding that bag-of-words performed worse than collocations. The accuracy of the collocation results are better than (Mohammad and Pedersen, 2004) unigram and bigram results. (Lee and Ng, 2002) evaluate unigrams and collocations on the Senseval-1 and Senseval-2LS data sets finding unigrams perform lower than collocations. The accuracy of the collocation results are also higher than the Senseval-1 and Senseval-2LS unigram and bigram results reported by (Mohammad and Pedersen, 2004). (Mohammad and Pedersen, 2004) found no difference in bigrams and unigrams for the Senseval-1 and Senseval-2LS data sets and marginal differences for the remaining data sets. (Joshi, Pedersen, and Maclin, 2005) feature set also includes unigrams and bigrams (although only the unigram results reported in this format). They use unigrams that surround the target word at the sentence level (s-unigrams) and abstract level (a-unigrams). (Joshi, Pedersen, and Maclin, 2005) found that the length of the sentences were too small to identify significant bigrams in most cases which may explain why there is very little difference in (Mohammad and Pedersen, 2004) unigram and bigram results.

(Lee and Ng, 2002) found using the POS of the surrounding words performs the same as the noun syntactic

relations (srelation) with an accuracy of 70%. (Ng and Lee, 1996) system returned an accuracy of 77% using the POS and morphology (M) features and 44% using the verb-object syntactic relation (VO relation). (Mohammad and Pedersen, 2004) found that the POS of the surrounding words obtain a higher accuracy than using only the POS of the head words and using the POS of the target word obtains lower accuracy than using whether or not the target word is a head word. (Mohammad and Pedersen, 2004) also found that using the POS of the word of the phrase containing the target word (POS of Phrase) and the POS of the head word of the target word's parent phrase (POS of Parent) return a higher accuracy of just using the word itself.

A feature vector may contain different types of features. (McRoy, 1992) was the first to use different types of features in the same feature vector. (Lee, Ng, and Chia, 2004) use a combination of POS of the surrounding words, unigrams, collocations and syntactic relations (srelation) reporting an accuracy of 65.6 on the Senseval-2LS and 72.4% on the Senseval-3LS data. (Mohammad and Pedersen, 2004) showed that certain pairs of features were redundant and others complementary. A feature pair is redundant if it classifies all the target words the same way and complementary if it does not. This is a crucial point when determining what features to use.

(Leroy and Rindflesch, 2004) evaluate their system on a subset of the NLM-WSD data set using the most frequent sense as the baseline (55%). The accuracy is cumulative in the order of the features on the table. The authors found a small improvement in accuracy over the baseline using whether the target word was a head word (Head) as a feature but found no substantial increase or decrease in performance when using the target word's POS. When adding the semantic types of the surrounding words that occur in the same sentences as the target word (ST sentence), the performance increases but when using the semantic types of the surrounding words in the same phrase (ST phrase) the accuracy decreased. This shows that the surrounding words at the phrase level do not contain enough information to disambiguate the target word. The *non-core* relations between the semantic types of surrounding words (NC relations) lowered the accuracy of the system. The *core* relations (C relations) did not but showed no improvement. The *core* and *non-core* relations between the semantic types of the target word and the semantic types of the surrounding words (C Sense Act. and NC Sense Act.) did not improve the accuracy of the system. (Joshi, Pedersen, and Maclin, 2005) show unigram feature performs better than the features used by (Leroy and Rindflesch, 2004).

(Liu, Teller, and Friedman, 2004) evaluate the performance of various algorithms using combinations of all the different lexical features based on the surrounding words and their orientation to the target word. The

results show the feature set containing all words within a window size of three and their orientation, and the three nearest two-word collocations performed best on the NLM-WSD subset. The feature set containing all words within a window size of three and their orientation, the three nearest two-word collocations and the surrounding words performed best on the general English data sets. The accuracy of the results reported by (Liu, Teller, and Friedman, 2004) are comparable to those reported by (Joshi, Pedersen, and Maclin, 2005). The results reported by (Leroy and Rindflesch, 2004) are not comparable to those reported by (Liu, Teller, and Friedman, 2004) because they use different subsets of the NLM-WSD data.

Table 3: Supervised WSD Results

		Senseval 1	Senseval 2LS	“line”	“hard”	“serve”	“interest”	NLM- WSD	Senseval 3LS
(Leacock, Towell, and Voorhees, 1993)	bag-of-words			76%					
(Mooney, 1996)	bag-of-words			72%					
(Pedersen, 2000)	bag-of-words			88%					
(Ng and Lee, 1996)	bag-of-words collocations POS + M VO relation						62% 80% 77% 77%		
(Mohammad and Pedersen, 2004)	unigrams bigrams POS Head Head of Parent POS of Phrase POS of Parent	67% 67% 68% 64% 60% 59% 58%	55% 55% 55% 52% 50% 53% 53%	75% 73% 60% 55% 60% 54% 54%	83% 89% 85% 88% 85% 82% 82%	73% 72% 76% 47% 57% 41% 42%	76% 79% 80% 69% 68% 55% 55%		
(Lee, Ng, and Chia, 2004)	POS+unigram +collocation +srelation		65.6%						72.4
(Lee and Ng, 2002)	unigrams collocations POS srelation	70% 74% 70% 70%	58% 61% 55% 55%						
(Joshi, Pedersen, and Maclin, 2005)	a-unigrams s-unigrams							73%/85% 76%/82%	
(Leroy and Rindflesch, 2004)	Head POS ST (phrase) ST (sentence) NC relation C relation NC Sense Act. C Sense Act.							58% 58% 61% 66% 55% 56% 60% 60%	
(Liu, Teller, and Friedman, 2004)	unigram + bigrams + orientation			90%	92%	92%	92%	87%	

3.2 Unsupervised WSD Systems

There are at most two components to an unsupervised WSD system (see Figure 3, Section 1): i) a module that generates a feature set for the input data (this does exist in all systems), and ii) the algorithm which determines the sense of the ambiguous words. We classify an unsupervised WSD system as knowledge-lean or knowledge-enhanced depending on the external data used by the system. A knowledge-lean unsupervised WSD system uses the input data, and an external corpus. The advantage of this system over a knowledge-dependent system is that it is language independent. A knowledge-dependent supervised WSD system uses the input data, an external corpus and is augmented with an external knowledge source such as machine readable dictionaries (MRD), and ontologies.

In this section, we discuss two types of knowledge-lean and knowledge-enhanced unsupervised WSD systems: clustering, and parallel-text-based. We then compare the results of the systems discussed in this section.

3.2.1 Clustering

Clustering algorithms groups similar instances of target words. There are three types of clustering algorithms used in the papers discussed in this section: agglomerative, divisive, and partitional algorithms. Agglomerative algorithms start with each feature vector in a separate cluster and merge clusters. Divisive methods start with all feature vectors in one cluster and split clusters and partitional algorithms divide the feature vectors into a preset number of clusters. A clustering algorithm requires a feature set containing a feature vector for each instance of the target word. The algorithm has two steps, first the feature vectors that share the same sense are clustered together using a clustering algorithm, and second, the clusters are labeled with the appropriate sense.

The problem that arises with clustering is that clusters do not always correspond with the number of senses and can produce very fine grained distinctions that are not in the sense inventory.

Table 4 shows the features and the papers of four knowledge-lean unsupervised WSD systems associated with those features. The features used are i) morphology, ii) part-of-speech (POS), iii) first-order co-occurrences, iv) first-order bigrams, v) second-order co-occurrences, vi) second-order bigrams, vii) un-

Table 4: Unsupervised WSD Features

		(Pedersen and Bruce, 1997)	(Pedersen and Bruce, 1998)	(Schütze, 1998)	(Purandare and Pedersen, 2004)
	Morphology	x	x		
	POS	x	x		
First-order	co-occurrences bigrams	x	x		x x
Second-order	co-occurrences bigrams			x	x x
Collocations	Unrestricted Content	x x	x x		

restricted collocations and viii) content collocations.

If the target word is a noun, Morphology is whether the noun is singular or plural. If the target word is a verb, morphology is the tense of the verb. POS is the part-of-speech of the target word and/or surrounding words. First-order co-occurrences are the number of times a word co-occurs within a specified position to the left or the right of the target word. First-order bigrams (Purandare and Pedersen, 2004) are similar only for co-occurrences, only while the position of the word does not matter it does for bigrams. The number of times “forward march” occurs is equal to the number of times “march forward” occurs. For bigrams, this is not necessarily the case. A frequency cutoff is used to include only those word pairs that are above a certain frequency. Second-order co-occurrences (Schütze, 1998) and bigrams (Purandare and Pedersen, 2004) utilize the first-order co-occurrences and bigrams respectively. They are highly frequent words that occur with the words in the first order co-occurrence matrix. Similarly with first-order co-occurrences and bigrams, the position of the position of the word does not matter for co-occurrences but does for bigrams. A collocation is the target word and a word occurring in a specified position to the left or the right of the target word. There are two types of collocations that are used: unrestricted collocations and content collocations. Unrestricted collocations use all the words within a specified position to the right and the left of the target word where content collocations only use the content words.

(Pedersen and Bruce, 1997) and (Pedersen and Bruce, 1998) use three feature sets: i) morphology, the POS of the word to the left and the two words to the right of the target word and the first-order co-occurrence of the 1st, 2nd and 3rd most frequent word, ii) morphology, unrestricted collocations with the two words

to the left of the target word and right of the target word and iii) morphology, the POS of the two words to the left and right of the target word and the content collocation of the word to the left and right of the target word. (Pedersen and Bruce, 1997) note that the feature sets used are better at distinguishing between noun senses than verb and adjective senses. (Purandare and Pedersen, 2004) evaluate first-order co-occurrences, second-order co-occurrences, first-order bigrams and second-order bigrams.

3.2.2 Parallel Text-based

A parallel text consists of a source language text and its translation in some target language. The two texts are said to be *sentence aligned* if it has been determined which sentences are translations of each other. A sentence in the source text and its corresponding sentence in the target text are called a *sentence pair*. The goal of such a system is to use the word translations to determine the sense of ambiguous words in the data set. An example of this is given by (Diab, 2000). The English word *bank* translates to the French word *banque*, to mean *financial institution* as well as the French words *rive* and *bord* to mean *shore line*. Therefore, if *bord* is seen as the translation of *bank* in the parallel text then the sense of the English word *bank* would be *shore line* rather than *financial institution*.

(Diab, 2000) introduces a knowledge-enhanced unsupervised WSD system that consists of four steps: i) sentence align the source and target data sets, ii) word align the data sets using a machine translation system, iii) using a taxonomy (such as WordNet) calculate the distance between the senses of the words in the target data set and assign a sense to each of the words, and iv) use the senses from the target data set to assign senses back to the source data set.

(Diab and Resnik, 2001) introduce a knowledge-lean unsupervised WSD system that consists of four steps: i) sentence align the source and target data sets, ii) group the words in the target data set that are translated to the the same word in the source data set, iii) label the group with a sense from the sense inventory, and iv) project these labels back to the source data set.

(Bhattacharya, Getoor, and Bengio, 2004) introduce two knowledge-enhanced unsupervised WSD systems: “Sense Model” and “Concept Model”. The Sense Model makes the assumption that the words in the target and source data sets are a one-to-one translation of each other. This means that a single word in the source data set translates to a single word in the target data set. The model predicts the sense of a word based

on Equation 1 which is the calculation for the joint probability that the word in the target data set, W_{TL} , and its translation in the source data set, W_{SL} , have the same sense, S . The possible senses for S and the distributions are obtained from the sense inventory, WordNet.

$$P(W_{TL}, W_{SL}, S) = P(W_{TL}|S)P(W_{SL}|S)P(S) \quad (1)$$

The Concept Model does not make a one-to-one assumption. It assumes that there is possibility that many words from the source data set could translate to many words in the target. The data is still a word-to-word translation but the the authors make the assumption that “the intended meaning of the target word does not always match perfectly with the intended meaning of the source word.”. The Concept Model predicts the senses of a word based on Equation 2 which is the calculation for the joint probability of the word in the word in the target data set, W_{TL} , and its translation in the source data set, W_{SL} , have the sense, S_{TL} and S_{SL} , and have the same concept, C . The possible senses for S are obtained from the sense inventory, WordNet. The distributions are obtained by “clustering the Spanish senses and then clustering the English and Spanish senses to build the concepts”. The words are clustered together based on how similar they are to each other. The authors use the similarity measure introduced by (Resnik, 1995) described in Section 3.3.

$$P(W_{TL}, W_{SL}, S_{TL}, S_{SL}, C) = P(W_{TL}|S_{TL})P(W_{SL}|S_{SL})P(S_{TL}|C)P(S_{SL}|C)P(C) \quad (2)$$

(Bhattacharya, Getoor, and Bengio, 2004) estimate the joint probabilities by using Sense and Concept Model with the Expectation Maximization algorithm (Dempster, Laird, and Rubin, 1977).

3.2.3 Unsupervised WSD Results (Accuracy)

Table 5 shows the results for the unsupervised WSD systems described above.

(Pedersen and Bruce, 1997) and (Pedersen and Bruce, 1998) evaluate their system using the “line” data and the (Bruce and Wiebe, 1994) corpus (*Mix1*). Their results both show that the feature set containing Morphology(M), POS of the surrounding words and the content collocations performed the best regardless

Table 5: Unsupervised WSD Results

Paper	features	<i>Mix1</i>	“line”	“hard”	“serve”	Senseval-2LS	SemCor	Senseval-2AW	<i>Mix2</i>
(Pedersen and Bruce, 1997)	M + POS + First-order Co-occurrence	65.5%							
	M + Unrestricted Collocations	65.3%							
	M + POS + Content Collocations	66.2%							
(Pedersen and Bruce, 1998)	M + POS + First-order Co-occurrence		64.6%						
	M + Unrestricted Collocations		65.7%						
	M + POS + Content Collocations		65.9%						
(Purandare and Pedersen, 2004)	First-order Collocations		62%	41%	37%	44%			
	First-order Bigrams		68%	87%	46%	44%			
	Second-order Collocations		55%	73%	34%	43%			
	Second-order Bigrams		38%	63%	31%	47%			
(Diab, 2000)	French (GP)						71.4%		
	German (GP)						69.1%		
	Spanish (GP)						70.5%		
	Spanish (SP)						70.6%		
	merged Spanish						76.9%		
(Diab and Resnik, 2001)	French (GP)							58.1%	
	Spanish (GP)							57.9%	
	French (SP)							58.0%	
	Spanish (SP)							60.0%	
	merged French							59.4%	
	merged Spanish							59.4%	
(Bhattacharya, Getoor, and Bengio, 2004)	Sense Model								62.5%
	Concept Model								67.2%
	(Diab and Resnik, 2001)								61.8%

of the clustering algorithm.

(Purandare and Pedersen, 2004) evaluate their system using the “line”, “hard”, “serve” and Senseval-2LS data sets. They show that first-order bigrams perform better on the “line”, “hard”, and “serve” data sets while second-order bigrams perform best on the Senseval-2LS data sets. (Purandare and Pedersen, 2004) note that Senseval-2LS is a smaller data set than the others which may have something to do with the difference in results. The authors also show that bigram features perform better than co-occurrence features regardless of the data set.

(Diab, 2000) evaluate their system using the SemCor data set and translated it into French, German and Spanish using Systran Professional 2.0 (SP) and Globalink Power Translator Prov v.6.4 (GP) machine translation (MT) packages. The results show that with 100% coverage an accuracy of 71.4%, 69.1% and 70.5% was achieved respectively for the French, German and Spanish translations done by GP MT package. An accuracy of 70.6% was achieved for the Spanish translations done by the SP MT package. An accuracy of 76.9% was achieved for a merging of the Spanish translations done by both systems.

(Diab and Resnik, 2001) evaluate their system using the Senseval-2AW data set and translated it into French and Spanish using Systran Professional 2.0 (SP) and Globalink Power Translator Prov v.6.4 (GP) machine translation (MT) packages. The results show an accuracy of 58.1% and 57.9% was achieved respectively for the French and Spanish translations done by GP MT package. An accuracy of 58.0 and 60.0% was achieved respectively for the French and Spanish translations done by the SP MT package. An accuracy 59.4% was achieved for both the French and Spanish when the translations were done by merging both translation systems output.

(Bhattacharya, Getoor, and Bengio, 2004) evaluate their system using the Senseval-1, Senseval-2AW, Senseval-2LS, the Brown Corpus and the Penn Treebank Wall Street Journals section 18-24 (*Mix2*). The authors translated the data set into Spanish using Systran Professional 2.0 (SP), Globalink Power Translator Prov v.6.4 (GP) and GIZA++ machine translation (MT) packages and the translations were combined. They compare their two systems with (Diab and Resnik, 2001). The results show an accuracy of 61.8%, 62.5% and 67.2% for the authors implementation of (Diab and Resnik, 2001), the Sense Model and the Concept Model respectively. (Cabezas, Bhattacharya, and Resnik, 2004) evaluate the Sense Model unsupervised WSD system introduced by (Bhattacharya, Getoor, and Bengio, 2004). They translate the Senseval-2AW data set into Chinese, Spanish and French using the GIZA++ MT package. (Cabezas, Bhattacharya, and Resnik, 2004)

results show an accuracy of 44.5%, 44.4%, and 44.5% for the Chinese, Spanish and French translations.

3.3 Knowledge-based WSD Systems

There is only one component in a knowledge-based WSD system (see Figure 2, Section 1). This component is the algorithm which determines the sense of the ambiguous words in the input data using external knowledge sources and potentially augmented by a corpus. Knowledge-based WSD system can be broken up into five subsets: i) similarity or relatedness-based, ii) contextual knowledge-based, iii) frequency-based, iv) graph-based, and v) vector-based.

3.3.1 Similarity and Relatedness-based

Here, we discuss 14 knowledge-based WSD systems that incorporate similarity and relatedness measures. Similarity and relatedness measures assign a score to how similar or related two concepts are to each other. In WSD, these are used to determine how similar or related the surrounding words of the target word are to the possible senses of the target word. These measures require the use of an ontology. The most commonly used ontology is WordNet. WordNet is a machine readable dictionary whose words are organized into concepts that connected together through a variety of relations. We distinguish between similarity and relatedness measures because concepts that are not similar can still be related. For example, *hot* and *cold* are related but not similar. Therefore, similarity score between them would be low while their relatedness score would be high.

In this section, we discuss 14 similarity and relatedness measures that have been used in WSD. We then discuss comparative analysis that have been made of these measures. The similarity measures are categorized as: path-based measures and information content (IC) measures. The path-based measures are: (Rada et al., 1989), (Sussna, 1993), (Leacock and Chodorow, 1998), (Altintas, Karsligil, and Coskun, 2005), (Wu and Palmer, 1994), (Agirre and Rigau, 1996), and (Hirst and St-Onge, 1998). The IC measures are: (Resnik, 1995), (J. Jiang, 1997) and (Lin, 1997). The relatedness measures are: (Lesk, 1986), (Cowie, Guthrie, and Guthrie, 1992), (Banerjee and Pedersen, 2003) and (Patwardhan, 2003).

Path-based Similarity Measures

Path-based measures are dependent on the length of the path between two concepts, c_1 and c_2 , in an ontology. (Rada et al., 1989) introduce the measure conceptual distance. Conceptual distance is calculated as the

shortest path between two concepts in an ontology. (Sussna, 1993) extended this measure by assuming that the shortest path from c_1 to c_2 may not be the same when going from c_2 to c_1 . The authors take an average of the shortest paths from both directions. (Leacock and Chodorow, 1998) (sim_{lch}) extend the conceptual distance measure by taking the negative log of the shortest path and dividing it by twice the total depth of the ontology (D) as defined in Equation 3. (Altintas, Karsligil, and Coskun, 2005) modified (Leacock and Chodorow, 1998) implementation by introducing a *SpecFactor* which takes into consideration the specificity of a concept using its location within a cluster. Concepts with close specificity values indicate higher similarity than those that are not.

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

(Wu and Palmer, 1994) (sim_{wup}) introduce a measure that takes into consideration the depth of two concepts in a ontology and the depth of their least common subsumer (LCS). The LCS is the most specific concept two concepts share as an ancestor. The measure is twice the LCS of two concepts is divided by the sum of their individual depths as defined in Equation 4.

$$sim_{wup}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \quad (4)$$

(Agirre and Rigau, 1996) (sim_{cd}) introduce the measure conceptual density. This measure takes into consideration both the depth of the individual concepts and the shortest path between them. This measure is defined in Equation 5 where $nhyp$ is the mean number of hyponyms per node, m is the number of senses of the target word and $descendants$ is the total number of words within the hierarchy of word c .

$$sim_{cd}(c_1, c_2) = \frac{\sum_{i=0}^{m-1} nhyp^i{}^{0.20}}{descendants_{c_1}} \quad (5)$$

(Hirst and St-Onge, 1998) (sim_{hirst}) introduce a measure that classifies the similarity between a pair of concepts as extra strong, strong, medium strong and weak. Two concepts are extra strong if their surface forms are identical. Two concepts are strong if one of the three following conditions are met: i) the path between them is horizontal, ii) one of the concepts can be represented by a compound word that contains the other concept, or iii) the path weight (sim_{hirst}) is at least some value $2 * C$. Two concepts are medium

strong if the path weight is at least C . The path weight is defined as some value C minus the length of the path ($pathLength$) minus the weighted number of changes in direction a path between two concepts as seen in Equation 6. (Budanitsky and Hirst, 2001) and (Pedersen, Banerjee, and Patwardhan, 2005) set C equal to eight and k equal to one this measure.

$$sim_{hirst}(c_1, c_2) = C - pathLength(c_1, c_2) - (k * theNumberofDirectionChanges) \quad (6)$$

Information Content Similarity Measures

Information content (IC) measures the specificity of a concept in a ontology. A concept with a high IC value is more specific to a specific topic than one with a low IC value. IC is formally defined as the negative log of the probability of a concept. The probability of a concept is calculated using a large corpora such as the British National Corpus.

(Resnik, 1995) modified information content to be used as a similarity measure. He defined the information content of two concepts to be the information content of their least common subsumer (LCS) as seen in Equation 7. As stated above the LCS is the most specific concept two concepts share as an ancestor.

$$sim_{res} = IC(lcs(c_1, c_2) = -\log(P(lcs(c_1, c_2))) \quad (7)$$

(J. Jiang, 1997) and (Lin, 1997) extended (Resnik, 1995)'s information content measure. (J. Jiang, 1997) modified it to include the length of the path between the two concepts as seen in Equation 8. (Lin, 1997) modified it to include the individual information content of the two concepts as seen in Equation 9.

$$sim_{jcn} = \frac{1}{IC(c_1) + IC(c_2) - 2 * IC(lcs(c_1, c_2))} \quad (8)$$

$$sim_{lin} = \frac{2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (9)$$

Relatedness Measures

The relatedness measures discussed in the literature are based on the overlap between the definitions (called glosses when using the MRD WordNet) of two concepts. An overlap is the longest sequence of one or more

consecutive words that occur in both definitions. These measures can be applied to WSD task by looking at the overlap between the words surrounding the target word and the gloss of the potential sense.

(Lesk, 1986) introduce a measure that determines the relatedness between two concepts by counting the number of overlaps between two glosses. There are two limitations to this measure: i) to calculate the overlap for all possible senses and all possible words is computationally infeasible and ii) the glosses are typically very short and therefore may not contain enough overlaps to distinguish between multiple concepts.

(Cowie, Guthrie, and Guthrie, 1992) introduce a measure called simulated annealing to alleviate the first limitation. They use the simulated annealing optimization algorithm to approximate the results of calculating all possible combinations of senses. Common words between the possible glosses of the target word and the definition of the surrounding words normalized based on the number of words in the definitions.

(Banerjee and Pedersen, 2003) introduce a measure to alleviate the second limitation by not only looking at the gloss of the concept but also the gloss of the related concepts. (Patwardhan, 2003) extends this approach further by representing the glosses (including the glosses of the related concepts) as vectors which we discuss in the vector-based knowledge-based WSD systems.

(Pedersen, 2004) use the measure introduced by (Banerjee and Pedersen, 2003) in his SenseRelate algorithm that obtains the relatedness score between the words surrounding the target word and each possible sense of the target word. The scores are summed for each sense and the target word is assigned the sense with the highest score.

Comparative Analysis of Semantic and Relatedness Measures

Table 6 shows the semantic and relatedness measures used by researchers in knowledge-based WSD. There are six papers and fourteen measures discussed in this section. Each paper compares at least two measures on various test sets. The order of the papers in the table is in the order they are discussed below.

(Pedersen et al., 2006) compare all of the path-based and information measures and the relatedness measures (Patwardhan, 2003) on a biomedical/clinical test set of 30 ambiguous words. The authors use SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terms) which is an ontology in the clinical domain the Mayo Clinic corpus of clinical notes to obtain the probabilities for the information content measures. They report (Patwardhan, 2003) outperforms all other algorithms.

Table 6: Semantic and Relatedness Measures used in Knowledge-based WSD Systems

	measure	(Pedersen et al., 2006)	(Pedersen, Banerjee, and Patwardhan, 2005)	(Altintas, Karsligil, and Coskun, 2005)	(Budanitsky and Hirst, 2001)	(Agirre and Rigau, 1996)	(Caviedes and Cimino, 2004)
Path-based Similarity Measures	(Leacock and Chodorow, 1998) (Wu and Palmer, 1994) (Hirst and St-Onge, 1998) (Agirre and Rigau, 1996) (Altintas, Karsligil, and Coskun, 2005) (Sussna, 1993) (Rada et al., 1989)	0.47	.23/.14 .30/.04 .20/.05	.32 .33 .35	.82/.84 .74/.79	.70 .65	.60-.77
Information Content Similarity Measures	(Resnik, 1995) (J. Jiang, 1997) (Lin, 1997)	0.55 0.55 0.69	.30/.05 .35/.12 .33/.06	.30 .31 .38	.77/.78 .85/.78 .83/.82		
Relatedness Measures	(Lesk, 1986) (Banerjee and Pedersen, 2003) (Cowie, Guthrie, and Guthrie, 1992) (Patwardhan, 2003)	0.76	.28/.13/.18 .41/.20/.23 .29/.19/.22				
	Metric	Correlation	F-measure	Precision	Correlation	Precision	Correlation
	Data Set	Clinical Data	Senseval-2LS	Senseval-2LS	R&G/M&C	Time Data	Biomedical Data

(Pedersen, Banerjee, and Patwardhan, 2005) compare the information content measures (Resnik, 1995) and the relatedness measures (Lesk, 1986), (Banerjee and Pedersen, 2003) and (Patwardhan, 2003). The authors use WordNet for their ontology and the British National Corpus and SemCor as their external corpus to obtain the probabilities for the information content similarity measures and relatedness measure (results reported above are using the British National Corpus). The authors run each of the measures on three tests sets comprising of nouns, verbs and adjectives from the Senseval-2LS data (we report the results as noun/verb and noun/verb/adjective when applicable). Since the adjectives in WordNet are not hierarchical, only the relatedness measures and (Hirst and St-Onge, 1998) can be used. The results show that for nouns and verbs, (Banerjee and Pedersen, 2003) significantly performs better than the other measures. For adjectives, (Patwardhan, 2003) and (Banerjee and Pedersen, 2003) outperform the other measures.

(Pedersen, Banerjee, and Patwardhan, 2005), (Altintas, Karsligil, and Coskun, 2005) and (Budanitsky and Hirst, 2001) compare the path-based similarity measures (Hirst and St-Onge, 1998), (Leacock and Chodorow, 1998), and (Resnik, 1995), the information content similarity measures (J. Jiang, 1997) and (Lin, 1997). (Budanitsky and Hirst, 2001) found using the correlation data from Rubenstein & Goodenough (R&G) and Miller & Charles (M&C). that (J. Jiang, 1997) and (Leacock and Chodorow, 1998) respectively outperformed the other measures. (Pedersen, Banerjee, and Patwardhan, 2005) results agree with (Budanitsky and Hirst, 2001) showing that (J. Jiang, 1997) performs the best out of the other measures using the Senseval-2LS data set. This is contrary to (Altintas, Karsligil, and Coskun, 2005) results who report (Lin, 1997) performs better using the same data set.

(Agirre and Rigau, 1996) show that their conceptual density measure performs better than (Sussna, 1993) from the Time Collection compiled by (Sussna, 1993). (Caviedes and Cimino, 2004) apply the conceptual distance measure (Rada et al., 1989) on the UMLS. They modified the measure by limiting the type of links that can be used when calculating the shortest path between two concepts in the UMLS.

3.3.2 Contextual Knowledge-based

Here, we discuss three knowledge-based WSD systems that incorporate contextual knowledge. The contextual knowledge used is: simulated annealing, selectional preferences, and subject codes.

(Yarowsky, 1992) introduce a knowledge-based WSD system that uses subject codes called categories from

the Roget Thesaurus. He calculates the mutual information of each of the surrounding words given the category of each of the possible senses of the target word. (Yarowsky, 1992) sums the log of this score for each word in the sentence. The sense with the greatest sum is chosen. The authors test their algorithm on sentences extracted from Grolier's Encyclopedia report and report a mean accuracy of 92% when determining the sense between three possibilities. (Yarowsky, 1992) notes the categories can also be obtained from WordNet or the Longman Dictionary of Contemporary English (LDOCE) subject codes.

(Stevenson and Wilks, 2001) introduce a knowledge-based WSD system that uses simulated annealing, selectional preferences and the subject codes derived from adapting (Yarowsky, 1992) algorithm from using Roget Thesaurus to LDOCE. Simulated annealing is discussed in Section 3.3.1. Selectional preferences are the preferred arguments that a specific verb takes. The selectional preferences are used to disambiguate nouns in the subject or object position by looking at the type of arguments that the verb takes. For example, "objects of the verb *eat* tend to be food items" (Manning and Schütze, 1999). The probability distributions for the selectional preferences are obtained using LDOCE. (Stevenson and Wilks, 2001) report that subject codes return a higher accuracy (79%) over simulated annealing (65%) and selectional preferences (45%) on the SemCor data. Combining the models using the exemplar based learning algorithm TiMBL, the authors report an accuracy of 94.65% at the homograph level and 90.37% at the sense level.

(McCarthy and Carroll, 2003) introduces a knowledge-based WSD system that uses selectional preference combined with the 'one-sense-per discourse' heuristic (Gale, Church, and Yarowsky, 1992a). The preferences they use are: subject, direct object, and adjective-noun. The preferences are obtained using the wide-coverage unification-based shallow parser developed by (Briscoe and Carroll, 1995). The authors use the preference model "Tree Cut Models" (TCM) (Li and Abe, 1998) to determine the sense of a target word. The distribution for the model is obtained from the British National Corpus. (McCarthy and Carroll, 2003) evaluated their system on the Senseval-2AW data set reporting a 51.1% accuracy when using the one-sense-per-discourse heuristic and 52.3% otherwise. The results are not directly comparable to (Stevenson and Wilks, 2001) but they show similar results.

3.3.3 Frequency-based

Here, we discuss two knowledge-based WSD systems that incorporate the frequency of the occurrence of the target word and its surrounding words.

(Mihalcea and Moldovan, 1999) introduce a knowledge-based WSD system that uses the frequency of word pairs to determine their appropriate sense. Given an untagged word pair (w_1, w_2) that occur next to each other in the sentence the system creates a list of similar words, s_i , for each sense of *one of the words* using WordNet, w_2 for example. The system then forms (s_i, w_2) pairs and queries the internet for each pair and ranks the pairs based on the number of documents returned by a search engine (the authors use AltaVista). This will rank the senses of w_2 . The same thing is done for word w_1 . If the (w_1, w_2) is not a verb-noun pair then the highest ranking sense is returned otherwise the *conceptual density* (Agirre and Rigau, 1996) (described above) is computed using Equation 10, where cc_{w_1, w_2} is the number of common concepts between w_1 and w_2 in WordNet, h_k is the level of noun k in the hierarchy of w_1 and $descendants_{w_2}$ is the total number times w_2 is in the hierarchy. (Mihalcea and Moldovan, 1999) evaluate their system on the SemCor data set reporting an overall average of 80.1%.

$$C_{w_1, w_2} = \frac{\sum_k^c c_{w_1, w_2} h_k}{\log descendants_{w_2}} \quad (10)$$

(Klapaftis and Manandhar, 2005) introduce a knowledge-based WSD system that uses WordNet to disambiguate a target word. WordNet concepts are connected through the relations *hypernym*, *hyponym*, *meronym*, *holonym* and *synonym*. (Klapaftis and Manandhar, 2005) create a list of words from topic signatures collected from Google by querying the sentence containing the target word. The words are separated into groups based on if they are one, two, or three links from the target word in WordNet when using a specific relation. This is done for each relation creating 15 groups. For each group, the authors store the sum of the words frequency in the groups corresponding tables (HyperFreq, HypoFreq, MeroFreq, HoloFreq and SynmFreq). (Klapaftis and Manandhar, 2005) then calculate a “target word score” by taking the average of the results returned from Equations using Equations 11, 12, and 13. This is repeated for each possible sense of the target word using the senses definition (gloss) from WordNet instead of the sentence. This creates a “sense score” for each possible sense. The target word is assigned the sense whose sense score is closest to the target word score. (Klapaftis and Manandhar, 2005) evaluate their system on the SemCor data set reporting an overall average of 58.9%.

$$Hyper = \frac{\sum_{i=1}^3 a_i HyperFreq[i]}{\sum_{i=1}^3 HyperFreq[i]} \quad (11)$$

$$Hypo = \frac{\sum_{i=1}^3 a_i HypoFreq[i]}{\sum_{i=1}^3 HypoFreq[i]} \quad (12)$$

$$MHS = \frac{a_1 MeroFreq + a_2 HoloFreq + a_3 SynmFreq}{MeroFreq + HoloFreq + SynmFreq} \quad (13)$$

Graph-based

Here, we discuss two graph-based knowledge-based WSD systems. The concept of a graph-based knowledge-based WSD system is that each node in the graph represents a sense of a specific word. The nodes are connected based on their probability of occurring together. The probabilities are then used to pick the most probable set of labels for a sequence of words.

(Mihalcea, 2005) introduce a knowledge-based WSD system that consists of three steps. In the first step, a label dependency graph is constructed by creating an edge between all the nodes in the graph and assigning the edge a “relatedness score”. The nodes consist of all possible senses of the target word and the sense of its surrounding words. The “relatedness scores” are obtained using the (Lesk, 1986) relatedness measure. (Lesk, 1986) calculates the relatedness between two nodes (senses) based on their definitions obtained from a machine readable dictionary (see Section 3.3. The authors use the dictionary the Longman Dictionary of Contemporary English (LDOCE). In the second step, the labels are assigned a “label score” using the graph-based ranking algorithm PageRank. In the last step, the target words are assigned a sense by identifying the node (sense) with the highest “label score”. To evaluate their system, (Mihalcea, 2005) use the Senseval-2AW, Senseval-3AW and a subset of SemCor. The author obtained a 55.3% accuracy on the Senseval-2AW data set, 52.2% on the Senseval-3AW data set, and 56.5% on the SemCor data set.

(Navigli and Velardi, 2005) introduce an iterative knowledge-based WSD system, SSI. The SSI system iteratively steps through a sentence assigning a sense to each content word. Initially, only words that are unambiguous are assigned a sense. In subsequent iterations, the target word is assigned a sense if there exists a relation between previously assigned senses and the possible senses for the target word. The authors used the machine readable dictionary WordNet whose concepts are connected through the relations *hypernym*, *hyponym*, *meronym*, *holonym* and *synonym*. The authors evaluate their system using the Senseval-3WN data set and obtained an accuracy of 86.0% for nouns, 69.4% for verbs and 78.6% for adjectives. This results in an overall accuracy of 78%.

3.3.4 Vector-based

Here, we discuss three vector-based knowledge-based WSD systems that use different knowledge sources: (Patwardhan, 2003) uses the MRD WordNet, (Humphrey et al., 2006) use the ontology UMLS, and (Mohammad and Hirst, 2006) use the MRT Macquarie's Thesaurus.

(Patwardhan, 2003) introduces a knowledge-based WSD system called "vector measure". A co-occurrence matrix containing the log likelihood between each word in a given machine readable dictionary and the words it co-occurs with in a given corpus is created. A co-occurrence matrix is a generic term for a matrix where the columns contain the information extracted from the knowledge source (in this case words from the dictionary) and the rows contain the words from the input data. A cell i in the matrix contains information about the co-occurrence of the column i -row i pair. In this case, the information is the log likelihood of the two words occurring together in a corpus. From the information in the co-occurrence matrix, a sense vector is created for each possible sense of a target word. The vector contains an average of the log likelihood between each of the words in the sense's definition and each of the words it co-occurs with in a specified corpus. Similarly, a target word vector is created containing an average of the log likelihood between the words surrounding the target word and the words they co-occur with in a corpus. The target word vector and its corresponding sense vectors are generated for each ambiguous word in input data. This set is passed to the algorithm which uses the cosine metric between the target word vector and each of its sense vectors. The sense of the vector with the smallest angle is assigned to the target word. This is done for each target word vector in the feature set.

(Pedersen, Banerjee, and Patwardhan, 2005) and (Pedersen et al., 2006) evaluate (Patwardhan, 2003) in the general English and clinical domain respectively. In the general English domain, they used the MRD WordNet. In the biomedical domain, they use SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terminology), an ontology covering the clinical domain. (The authors compared their results to several knowledge-based systems - see Similarity and Relatedness Measures in this section.)

(Humphrey et al., 2006) introduce a knowledge-based WSD system to be used to disambiguate words in MEDLINE journals by MetaMap. MetaMap is a program that maps biomedical text to concepts in the UMLS Metathesaurus (see Section 5.1 and 2.3.1). MetaMap currently uses a partial matching program, Medical Text Indexor (MTI), to determine the appropriate concept of a word. (Humphrey et al., 2006)

propose an system called “Journal Descriptor Indexing (JDI) of semantic type (ST) documents”. JDI is a ranking algorithm that automatically assigns Journal Descriptors (JD) to journal titles in order to maintain a subject index of all journals in MEDLINE. JDs are terms that are assigned to a journal that describe the type of articles it contains such as “Cardiology” and “Surgery”. The authors apply the JDI algorithm to semantic types. A semantic type is a categorization of a concept for example, *organism*. A concept may have more than one semantic type. A ST document is a set of Metathesaurus words associated with a particular semantic type in the Semantic Network. This is created by extracting single word Metathesaurus terms that are assigned a specific semantic type. An ST vector for each semantic type is created containing its “Majority Percentage”. The “Majority Percentage” (measure) is the number of times the word occurs in “phrases corresponding to UMLS concepts assigned that specific semantic type” (Humphrey, Rindfleisch, and Aronson, 2000) divided by the number of times the word occurs in “phrases corresponding to UMLS concepts” (Humphrey, Rindfleisch, and Aronson, 2000). A target word vector is created using the words surrounding the target word instead of the Metathesaurus terms. As in (Patwardhan, 2003), the target word vector and its corresponding sense vectors are generated for each ambiguous word in input data. This set is passed to the algorithm which uses the cosine metric between each of the ST vectors and the target word vector is calculated. This is done for each target word vector in the feature set. The main problem with this system is that two possible concepts may have the same semantic type(s) which would not allow for this system to work. The authors evaluate their system using a subset of the NLM-WSD data set and achieve an overall accuracy of 68.26%. This subset is the same used by (Leroy and Rindfleisch, 2004) who obtained a 65.6% accuracy evaluating their supervised WSD system (see Section 3.1).

(Mohammad and Hirst, 2006) introduce a knowledge-based WSD system that introduces the concept of a *Dominance* metric. They create a co-occurrence matrix containing the number of times a word occurs within a specified distance of a category in a given corpus. A category is obtained from the machine readable thesaurus, *Macquarie*, and the frequency information is obtained from the British National Corpus. The co-occurrence information is used to calculate the association between the words in the vocabulary and the thesaurus categories (senses). The authors analyze four measures of association Dice, Cosine, Pointwise Mutual Information (pmi), Odds Ratio (odds), Yule’s coefficient of colligation (yule) and Phi Coefficient (ϕ). The use of measures of association is similar to (Patwardhan, 2003) who use the log likelihood association measure. A feature vector is created for each possible category (sense) of a target word. The vector contains the association of the words surrounding the target word the category. This is stored in the input data feature

set.

The *Dominance* of a category is calculated based on the association scores from the vectors stored in the feature set. The category with the highest *Dominance* score is assigned to the target word. The authors give four different equations that can be used to calculate the *Dominance*: i) the normalized sum of the association scores between the surrounding words and the category; ii) the maximum association of the surrounding words and the category divided by the number of surrounding words; iii) the normalized sum of the association scores between all possible surrounding words of a target word and the category (all possible surrounding words is a union of all surrounding words co-occurring with in a specific distance of the target word in the entire corpus not just in the sentence); and iv) the maximum association of all possible surrounding words of a target and the category.

(Mohammad and Hirst, 2006) evaluate their system on the Senseval-1 data set. They did not give overall accuracy results for the entire dataset. They instead reported the accuracy of word groupings based on the number of possible senses, for example, words that contain only one sense are grouped together, those that contain two are grouped together. The authors showed that *Dominance* calculations one and three outperform calculations two and three. The odds ratio, pmi, and Yule's coefficients did not return significantly different results and significantly outperform the dice coefficient and cosine.

3.3.5 Knowledge-based WSD Results

Table 7 shows the results for the knowledge-based WSD systems described above. We only compare the relatedness-based systems introduced by (Banerjee and Pedersen, 2003) due to its performance and comparability with other systems. A more indepth analysis of the different measures is discussed above.

The contextual knowledge-based system introduced by (Stevenson and Wilks, 2001) performs with a 90.37% accuracy on the SemCor data set. The authors use a combination of subject codes, simulated annealing and selectional preference. The system is comparable to the frequency-based systems introduced by (Mihalcea and Moldovan, 1999) and (Klapaftis and Manandhar, 2005) that perform with an 80.1% and 58.9% accuracy respectively.

The graph-based system introduced by (Mihalcea, 2005) performs with a 55.3% accuracy on the Senseval-2AW data set. This system is comparable to the contextual knowledge-based system introduced by (Mc-

Table 7: Knowledge-based WSD Results

	Contextual Knowledge			Frequency-based		Graph-based		Vector-based		Relatedness-based	
	(Yarowsky, 1992)	(Stevenson and Wilks, 2001)	(McCarthy and Carroll, 2003)	(Mihalcea and Moldovan, 1999)	(Klapaftis and Manandhar, 2005)	(Mihalcea, 2005)	(Navigli and Velardi, 2005)	(Patwardhan, 2003)	(Humphrey et al., 2006)	(Banerjee and Pedersen, 2003)	(Pedersen, 2004)
subset Groiler's	92										
SemCor		90.37		80.1	58.9						
subset SemCor						56.5					
Senseval-2AW			52.3			55.3					
Senseval-3AW						52.2					
Senseval-3WN							78.0				
Senseval-2LS								23.3		28.0	
Senseval-3LS											40.3
NLM-WSD									68.3		
Metric	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	F-measure	Accuracy	F-measure	Accuracy

Carthy and Carroll, 2003) that perform with 52.3% accuracy. (Mihalcea, 2005) also evaluated their system on the Senseval-3AW data set and report a 52.2% accuracy.

The relatedness-based system introduced by (Banerjee and Pedersen, 2003) performs with a 28.0% F-measure on the Senseval-2LS data set. This is comparable to the vector-based measure introduced by (Patwardhan, 2003) that performs with a 23.3% F-measure. The relatedness-based system introduced by (Pedersen, 2004) reports an accuracy of 40.3% on the Senseval-3LS data set. This system uses the relatedness measure introduced by (Banerjee and Pedersen, 2003).

3.4 Bootstrapping WSD Systems

A bootstrapping WSD system automatically creates sense-tagged training data using an unsupervised WSD system for a supervised WSD systems (see Figure 5, Section 1).

(Yarowsky, 1995) introduce a bootstrapping WSD system that uses two heuristics: i) “one sense per collocation” and ii) “one sense per discourse. One sense per collocation is based on the assumption that “nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship. “One sense per discourse” is based on the assumption that as the sense of a “word is highly consistent within any given document”. Their system as five steps: i) identify all instances of ambiguous words in a corpus, ii) for each possible sense of each target word in the corpus, identify a small set of training example that are representative of that sense (this step can be done manually), iii) train the supervised WSD system’s algorithm using the small set of training examples, iv) run the supervised WSD on the remainder of the corpus applying the one sense per collocation and discourse heuristics, v) add those instances that are tagged with a high confidence to the seed set, and vi) repeat until all words have been tagged with high confidence. This final set is then used to train to retrain the supervised WSD system and tag the input data. (Yarowsky, 1995) evaluates his system from a data set extracted from a 460 million word corpus containing news articles, scientific abstracts, spoken dialog and novels. They report an average performance of 97% accuracy.

(Mihalcea, 2002) introduces a bootstrapping WSD system that consists of six steps: i) create a seed set from sense-tagged instances in SemCor, example sentences from WordNet and data from (Mihalcea and Moldovan, 1999), ii) for each sentence in the seed set query the web and extract sentences that contain the target word associated with the queried sentence, iv) perform WSD on the extracted sentences using the seed set as training data using (Mihalcea and Moldovan, 2000) knowledge-based WSD system, and v) add the sense-tagged extracted sentences to the seed set and repeat. (Mihalcea, 2002) evaluates her system on the target words “art”, “chair”, “church”, “detention” and “nation” from the Senseval-2LS and Senseval-2AW data set reporting an overall accuracy of 74.3%.

Table 8: Qualitative Analysis of WSD Systems

		Senseval 2LS	Senseval 2AW	Senseval 3LS	Senseval 3AW
Supervised WSD systems	(Lee and Ng, 2002)	74.0%			
	(Lee, Ng, and Chia, 2004)	65.6%		72.4%	
Bootstrapping WSD systems	(Mihalcea, 2002)		74.3%		
Unsupervised WSD systems	(Diab and Resnik, 2001)		60.0%		
	(Bhattacharya, Getoor, and Bengio, 2004)		44.5%		
	(Purandare and Pedersen, 2004)	47.0%			
Knowledge-based WSD systems	(McCarthy and Carroll, 2003)		52.3%		
	(Pedersen, 2004)			40.3%	
	(Mihalcea, 2005)		55.3%		52.2%

3.5 Overview of WSD Systems

Table 8 shows the accuracy of nine WSD systems discussed in this section. The results show that the problem with unsupervised and knowledge-based WSD systems is that they do not perform as well as state-of-the-art supervised and bootstrapping systems. The problem with supervised and bootstrapping WSD systems is that they are “less scalable than unsupervised methods because they rely on training data which may be costly and unrealistic to produce, and even then might be available for only a few ambiguous terms” (Widdows et al., 2003). Figure 6 shows a qualitative analysis between the accuracy and scalability of the best scoring WSD systems in Table 8.

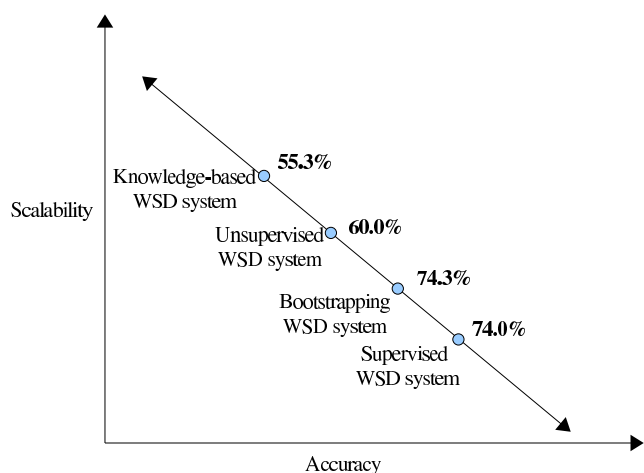


Figure 6: Qualitative Analysis of WSD Systems

4 Preliminary Work

In this section, we discuss three research questions that are addressed by our preliminary work.

(Joshi, Pedersen, and Maclin, 2005) use the unigram feature to disambiguate words in biomedical text. The unigram feature consists of words surrounding the target word with a high frequency while (Leroy and Rindflesch, 2004) use the semantic type of the UMLS Concept of the surrounding words but not the concepts themselves. This lead us to our first question: would using the UMLS Concepts of the surrounding words be an improvement over using the unigrams or semantic types of the surrounding words?

(Humphrey et al., 2006) introduce an unsupervised WSD system to disambiguate words in biomedical text that utilize Journal Descriptor Indexing. A Journal Descriptor is a manually assigned tag to journals in the National Library of Medicines' List of Serials Indexed for Online Users. (Gale, Church, and Yarowsky, 1992a) 'one-sense-per discourse' heuristic states that all instances of a target word have the same sense in a single discourse. This lead us to our second question: would the Journal Descriptors themselves contain enough information to disambiguate words in biomedical text using a supervised WSD system?

(Joshi, Pedersen, and Maclin, 2005) use the SVM algorithm in their supervised WSD system while (Leroy and Rindflesch, 2004) use the Naive Bayes algorithm in their system. (Joshi, Pedersen, and Maclin, 2005) report better results than (Leroy and Rindflesch, 2004) but the algorithms and the feature set are different. This leads us to our third question: would changing the algorithm change the accuracy of the results if the feature set remained the same?

4.1 UMLS Concepts

A UMLS Concept is defined as the "meaning" of a word in the UMLS. It is one level of abstraction away from the word itself. A given word may have multiple concepts. For example, the word "mole" could mean the mammal or a benign melanocytic nevus of the skin. Similarly multiple words can have the same concept. For example, "exocytosis" and "secretion" both can mean the cellular release of material within membrane-limited vesicles by fusion of the vesicles with the cell membrane.

The UMLS Concepts of the words surrounding the target word are annotated in the NLM-WSD data set. This data is in NLM-format. We converted the NLM data into XML format for easier readability and then

converted it to ARFF format required by the WEKA data-mining package (Witten and Frank, 1999).

4.2 Journal Descriptors

A Journal Descriptor is a manually assigned tag to the journals in the National Library of Medicines' List of Serials Indexed for Online Users. The journals in the Serials include a majority of journals indexed in MEDLINE which is where the NLM-WSD data originates. A journal may be given more than one descriptor. All citations inherit the Journal Descriptors of their respective journals. There currently exists 123 Journal Descriptors; examples of these are "Cardiology" and "Surgery".

The Journal Descriptors are not annotated in the NLM-WSD data set. We obtained the Journal Descriptors for the instances in NLM-WSD data set, using two step process. First, we wrote software to automatically query PubMed with the instances PubMed ID (PMID) that can be found in the NLM-WSD data set. PubMed is a National Library of Medicine search engine that contains over 16 million citations from MEDLINE and other biomedical articles dating back to the 1950s. We extracted the journal that contained the instance from the query results. Second, we obtained the Journal Descriptors that correspond to the journal from the "The List of Serials Indexed for Online Users, 2006" we be downloaded from www.nlm.nih.gov/tsd/serials/lsiou.html.

4.3 Results and Discussion

In this section, we first discuss the results of using the UMLS Concepts and Journal Descriptors as features. We compared our results to those reported by the literature when applicable and a "majority sense" baseline which is the accuracy that would be achieved if all the instances were classified to the sense with the greatest number of instances. Second, we analyze the results of using the Naive Bayes and SVMs algorithms.

4.3.1 Surrounding UMLS Concepts

To address our first question, we compared the results obtained using the UMLS Concept feature with the majority sense baseline and the best results obtained by (Leroy and Rindflesch, 2004) and (Joshi, Pedersen, and Maclin, 2005) on the NLM-WSD data set using the same algorithm used by the individual authors.

(Leroy and Rindflesch, 2004) used the Naive Bayes algorithm from the WEKA datamining package (Witten and Frank, 1999) and (Joshi, Pedersen, and Maclin, 2005) used the Support Vector Machine, SMO, from the same package.

(Leroy and Rindflesch, 2004) best reported results are a combination of the following features: i) whether the target word is the head word, ii) the POS of the head word, and iii) the semantic type of the words in the same sentence as the target word. We refer to this combination henceforth as “LeroyR04”.

Table 9 shows a comparison between the baseline, LeroyR04, the UMLS Concept (Concept) and a combination of Concept and LeroyR04 (Combo). The baseline is a “majority sense” baseline which is the accuracy that would be achieved if all the instances were assigned to the sense with the greatest number of instances. The overall average results show that overall Concept increases the accuracy of LeroyR04 by 5% and Combo increases it by 8%. They also show unlike LeroyR04, Concept and Combo significantly ($p \leq .05$ and $p \leq .01$ respectively) improve the baseline results by approximately 15% and 18% respectively.

Table 10 shows a comparison between the baseline, the unigram feature using the surrounding unigram at the sentence level (s-unigrams) and abstract level (a-unigrams) (Joshi, Pedersen, and Maclin, 2005), the UMLS Concept feature (Concept), the UMLS semantic type feature (ST) and the “bag-of-words” feature.

The unigram feature are the content words surrounding the target word that occur a specified number of times in a specified window. The surrounding concepts (Concept) and the surrounding semantic types (ST) used by (Leroy and Rindflesch, 2004) can be viewed as consecutively abstracting away from the content words in the sentence. The unigram feature uses the words themselves, the Concept uses the concepts of the words and ST uses the semantic type of the concepts of the words. Table 10 show Concept improves on the overall baseline results by approximately 13% but does not improve upon the overall results reported by (Joshi, Pedersen, and Maclin, 2005). Although the decrease is not significant, it is possible that the farther abstracted away a feature gets from the individual word itself, the lower the accuracy obtained. If this was the case, we would expect that using only the surrounding words in the sentence (“bag-of-words”) would achieve better results than Concept. This is not the case though, “bag-of-words” does not achieve a higher accuracy than concept. Therefore, the frequency information used in the unigram feature may be important to the disambiguation process.

Table 9: UMLS Concept Results using Naive Bayes

target word	baseline	LeroyR04	Concept	Combo
adjustment	62	57	74	74
blood pressure	54	46	60	59
degree	63	68	74	74
evaluation	50	57	59	62
growth	63	62	63	66
immunosupression	59	63	64	62
man	58	80	86	86
mosaic	52	66	67	75
nutrition	45	48	60	61
radiation	61	72	75	78
repair	52	81	74	78
scale	65	84	80	84
sensitivity	48	70	73	79
weight	47	68	67	79
white	49	62	76	82
<i>overall average</i>	55	65.6	70.13	73.27
versus baseline	–	+10%	+15% ($p \leq .05$)	+18% ($p \leq .01$)
versus LeroyRO4	–	–	+5%	+8%
versus Concept	–	–	–	+3%

Table 10: UMLS Concept Results using SVMs

target word	baseline	s-unigram	a-unigrams	Concept	ST	bag-of-words
adjustment	62	70	71	72	56	72
blood pressure	54	62	53	53	49	52
degree	63	92	89	73	70	75
evaluation	50	62	69	63	47	65
growth	63	63	71	60	55	65
immunosuppression	59	72	60	63	57	58
man	58	92	89	85	61	86
mosaic	52	77	87	71	72	73
nutrition	45	63	52	52	51	41
radiation	61	69	82	70	66	60
repair	52	72	87	67	69	68
scale	65	80	81	79	63	79
sensitivity	48	76	88	73	65	76
weight	47	80	83	71	52	62
white	49	72	79	72	62	65
<i>overall average</i>	55	73.47	76.07	68.27	59.47	66.47
versus baseline	–	+18% ($p \leq .01$)	+21% ($p \leq .01$)	+13%	+4%	+11% (
versus s-unigram	–	–	+3%	-5%	-14% ($p \leq .05$)	-6%
versus a-unigram	–	–	–	-8%	-17% ($p \leq .025$)	-9%
versus Concept	–	–	–	–	+9%	+2%
versus ST	–	–	–	–	–	+7%
versus bag-of-words	–	–	–	–	–	–

4.3.2 Journal Descriptors

To address our second question, we compared the results obtained using the Journal Descriptor feature with the majority sense baseline. Table 11 shows that the baseline outperforms the Journal Descriptor (JD) feature by approximately 4%. A closer analysis of the data shows that only 141 journals in 1970 were annotated with a Journal Descriptor. Running the algorithm on only those instances that have a Journal Descriptor (JD-subset) obtains an accuracy approximately 31% above the baseline. Analysis of the subset though showed that a majority of instances that contain Journal Descriptors have the same sense. For example, the target word “nutrition” contains 46 instances of the first sense, 17 of the second, 39 of the third and 12 of *none* of the sense. Only 45 of the first sense, one of the second, three of the third and zero of none of the senses have Journal Descriptors associated with them.

Table 11: Journal Descriptor Results using Naive Bayes

target word	baseline	JD	JD-subset
adjustment	62	63	67
blood pressure	54	61	87
degree	63	68	92
evaluation	50	56	95
growth	63	39	95
immunosuppression	59	60	98
man	58	59	98
mosaic	52	56	98
nutrition	45	49	92
radiation	61	62	98
repair	52	54	96
scale	65	9	-
sensitivity	48	50	98
weight	47	34	85
white	49	51	86
<i>overall average</i>	55.2	51.4	86.4

4.3.3 Algorithm Results

To address our third question, we compared the Naive Bayes and SVM algorithms on the three types of features sets, lexical, syntactic and semantic. Table 12 shows the performance of the features from each of the three feature sets for the Naive Bayes and SVM algorithms.

We compared the Naive Bayes and SVM using the “bag-of-words” lexical feature. This feature is used by (Leacock, Towell, and Voorhees, 1993), (Mooney, 1996) and (Pedersen, 2000) in their supervised WSD systems. Our results show that there is not a significant difference between the accuracy returned by the Naive Bayes and the SMV.

We compared the Naive Bayes and SVM using two syntactic features: the head word of the target word, its part-of-speech (POS), and a combination of the two. The individual overall results show that Naive Bayes and SVM assigns the data with approximately the same accuracy when using the head information and the POS information. When both sets of information was given, the Naive Bayes performed approximately 1% better than the SVM.

We compared the Naive Bayes and SVM using two semantic features: the UMLS Concepts (Concept) of the surrounding words, the semantic type (ST) of the surrounding words and a combination of the two (ST-C). The individual overall results and the combined overall results show that the Naive Bayes consistently reports a higher accuracy than the SVM.

4.4 Preliminary Conclusion

We addressed three research questions in our preliminary work: i) would using the UMLS Concepts of the surrounding word be an improvement over using the unigrams or semantic types of the surrounding words? ii) would the Journal Descriptors themselves contain enough information to disambiguate words in biomedical text using a supervised WSD system? and iii) would changing the algorithm change the accuracy of the results if the feature set remained the same?

In addressing the first question, we introduce a new feature UMLS Concept that can be used for the WSD task Table 13 shows the overall results for the baseline, (Joshi, Pedersen, and Maclin, 2005), (Leroy and Rindflesch, 2004),the UMLS Concept and a combination of Concept and (Leroy and Rindflesch, 2004)

Table 12: Supervised WSD Algorithm Results using Naive Bayes and SVM

	Lexical Features		Syntactic Features						Semantic Features					
target word	Naive Bayes	SVM	Naive Bayes			SVM			Naive Bayes			SVM		
	bag-of-words	bag-of-words	head	POS	head-POS	head	POS	head-POS	ST	Concept	ST-C	ST	Concept	ST-C
adjustment	71	72	62	62	62	62	62	62	53	74	74	56	72	70
blood pressure	60	52	54	54	51	51	54	52	54	60	59	46	53	53
degree	73	75	66	63	66	67	63	67	64	74	74	70	73	76
evaluation	59	65	50	50	50	49	50	49	50	59	61	47	63	59
growth	68	65	63	63	63	63	63	63	55	63	63	55	60	59
immunosuppression	68	58	57	59	57	59	59	59	62	64	62	57	63	63
man	84	86	62	60	58	55	60	57	63	86	86	61	85	86
mosaic	79	73	52	52	46	52	52	52	75	67	73	72	71	73
nutrition	53	41	45	45	53	45	45	44	55	60	61	51	52	53
radiation	73	60	61	61	61	61	61	61	66	75	78	66	70	67
repair	74	68	57	52	58	58	52	57	68	74	79	69	67	67
scale	81	79	79	67	79	78	67	77	70	80	80	63	79	80
sensitivity	72	76	54	48	54	49	48	49	64	73	80	65	73	74
weight	71	62	66	61	66	54	61	59	67	67	76	52	71	68
white	63	65	49	53	49	49	53	53	63	76	82	62	72	78
<i>overall average</i>	69.93	66.47	58.47	56.67	58.20	56.80	56.67	57.40	61.93	70.13	72.53	59.47	68.27	68.4

(LeroyR04). We showed that the surrounding UMLS concept increases results reported by (Leroy and Rindfleisch, 2004) (Leroy04) by 8% and significantly increases the baseline results by 15%. This comparison was done using the Naive Bayes algorithm because that is the algorithm used by (Leroy and Rindfleisch, 2004) for their experiments. The results though were lower than the unigram feature results reported by (Joshi, Pedersen, and Maclin, 2005) but not significantly. This comparison was done using the SVM algorithm because that is the algorithm used by (Joshi, Pedersen, and Maclin, 2005) for their experiments. The unigram feature takes into account how often a surrounding word is seen with the target word which improves the bag-of-words approach. We believe that a similar approach may help improve the UMLS Concept results.

In addressing the second question, we introduce a new feature Journal Descriptors. We showed that using the Journal Descriptor as features has potential but can not say how much due to the limited number of journals in our data set that have a Journal Descriptors.

Table 13: Overall Feature Results

feature	Naive Bayes	SVM
baseline	55	55
(Joshi, Pedersen, and Maclin, 2005)	–	76.07
LeroyR04	65.60	–
Concept	70.13	68.27
Concept + LeroyR04	73.27	68.60

In addressing the third question, we compared two commonly used Supervised WSD algorithms, Naive Bayes and SVM, using feature from three different categories: lexical features, syntactic features and semantic features. Table 14 shows the overall results of these experiments and their p – value. We found that irregardless of the feature set neither algorithm performed significantly better than the other.

Table 14: Overall Algorithm Results

feature set	Naive Bayes	SVM	p-value
lexical	69.93	66.47	> 1
syntactic	58.20	57.40	> 1
semantic	72.53	68.40	> 1

5 Proposed Work

The thesis of the proposed work is to investigate techniques to create a scaleable WSD system that maintains acceptable accuracy. In order to address this, we need a scalable WSD system that returns reasonable accuracy. Figure 7 shows a graph of the scalability versus accuracy of WSD systems. To date, state-of-the-art a supervised WSD system returns the highest accuracy but manual effort is needed to get sufficient training data for each word that needs to be disambiguated making them intractable for large scale problems such as information retrieval. Unsupervised and knowledge-based WSD systems do not require training data allowing them to scale up for large scale problems but return lower accuracy.

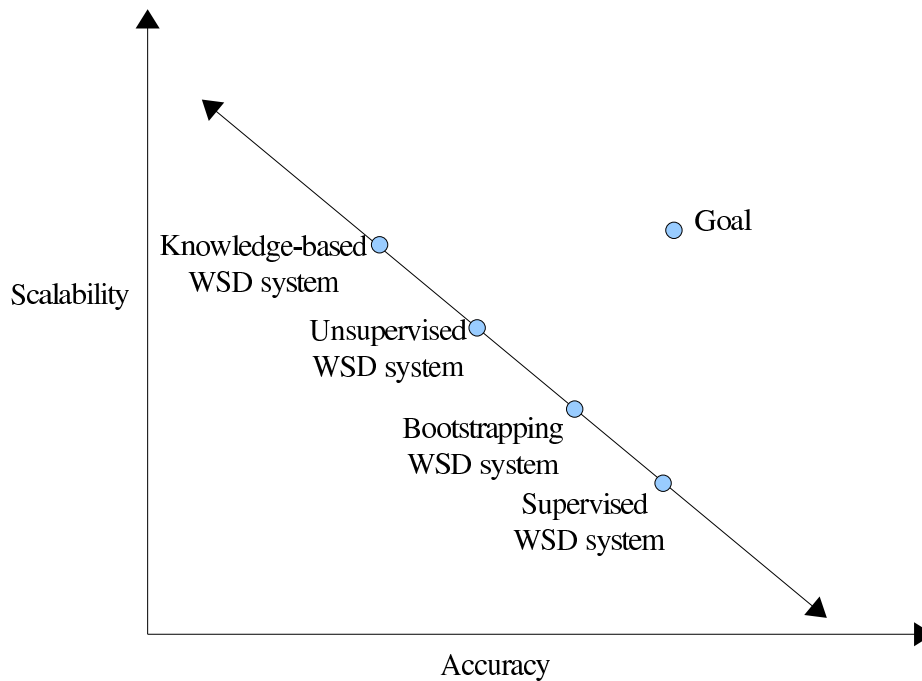


Figure 7: Scalability versus Accuracy of WSD systems

Our goal is to create a WSD system that is **both** scalable and achieves accuracy that is comparable to a state-of-the-art supervised WSD system. The thesis leads us to a specific question: can WSD on Medline abstracts be improved by incorporating knowledge from the UMLS?

Performing WSD on Medline abstracts is a large problem. Medline is an online database that contains 11 million references to biomedical journal articles and adds approximately 8,000 new articles every week. The words in each Medline abstract is mapped to a UMLS concept by the concept mapping system, MetaMap.

This allows researchers to query Medline for articles that contain specific concepts aiding them in finding relevant articles. In order to ensure a contents word is mapped to the appropriate concept, a scalable WSD system is required. For example, the word “growth” has three concepts in the UMLS: “Growth Aspects”, “Growth”, and “Tissue Growth”. Currently, MetaMap does not incorporate such a system. They map a word to a concept through a pattern matching algorithm using lexical variations of the words.

Figure 8 shows a diagram of our proposed knowledge-based WSD system. The system takes in Medline abstract as input data and assigns each word in the abstract a sense (concept) from the sense inventory UMLS. The words in the input data are disambiguated using information from biomedical articles and the UMLS.

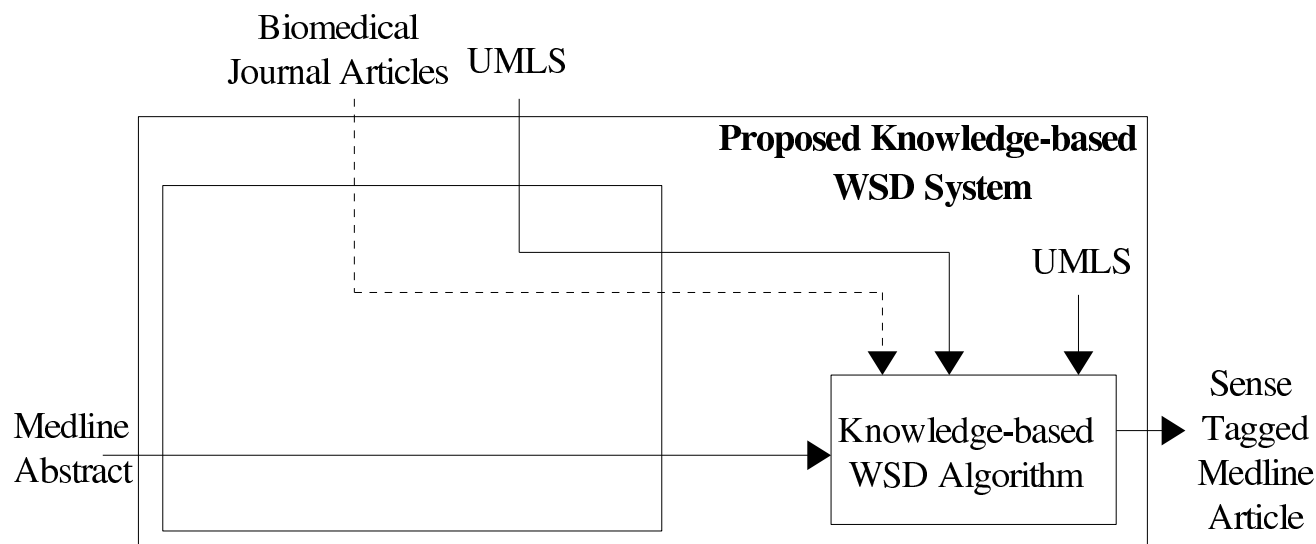


Figure 8: Proposed Knowledge-based WSD System

In the following sections, we discuss MetaMap in more detail and then propose a WSD-enhanced MetaMap system that incorporates our knowledge-based WSD system. We then discuss how we plan to evaluate our knowledge-based WSD system and the contributions of such a system. Lastly, we discuss various ways forward.

5.1 MetaMap

MetaMap was developed to improve retrieval of biomedical articles such as MEDLINE citations. To date, MetaMap does not have a WSD component in its system. (Aronson, 2001) notes that a WSD component would greatly improve the accuracy of the system's mappings.

Figure 9 shows the MetaMap system. It has five components: the preprocessor, the lexical variant generation (LVG) module, the candidate retrieval module, the candidate evaluation module and the mapping construction module.

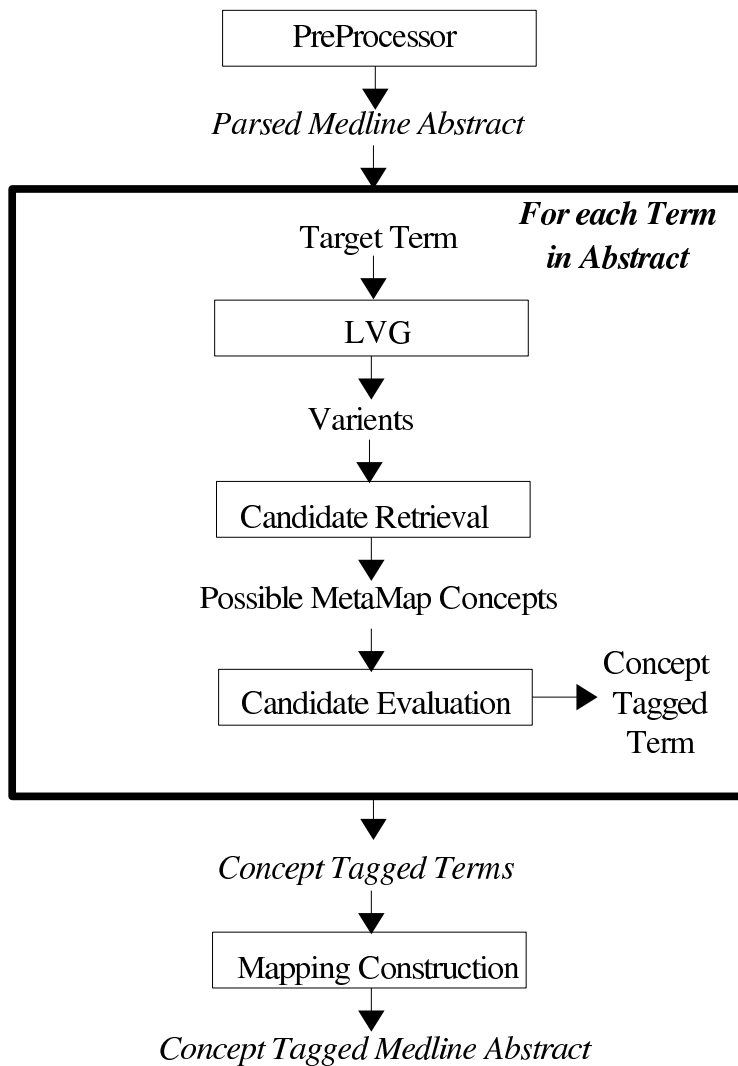


Figure 9: Current MetaMap System

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 “Medical Text Indexer” (MTI) 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. (Aronson, 2001) do not specifically state what metric is used except “an average of inverse distance scores”. 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 MTI score is assigned to the associated term. The mapping construction module generates the “Concept Mapped Medline Abstract”.

5.2 WSD-enhanced MetaMap System

Figure 10 shows a possible MetaMap system embedded with our proposed knowledge-based WSD system. The difference between the two systems is in the *Candidate Evaluation* module. This module is removed and our proposed knowledge-based WSD system is put in its place. The parsed Medline abstract and the target word are read into the knowledge-based WSD system. The system determines the appropriate UMLS concept (sense) of the word based on information from an external corpus of Medline articles and the UMLS using the *Possible MetaMap Concepts* as the sense inventory.

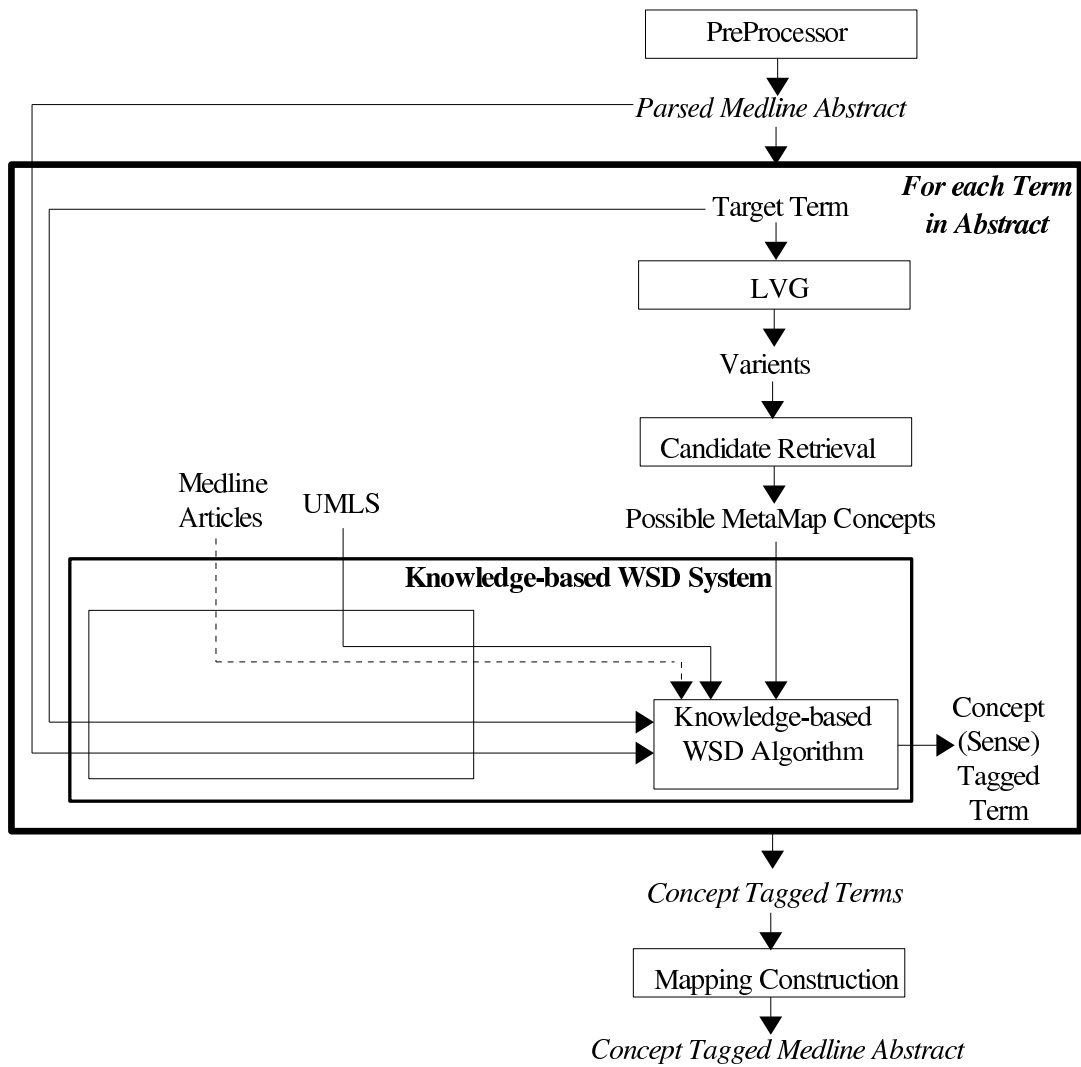


Figure 10: WSD-enhanced MetaMap System

5.3 Evaluation of Proposed Knowledge-based WSD System

To determine how well our proposed WSD system performs, we have a three part evaluation plan.

The first part is to evaluate our system against MetaMap using the NLM-WSD data set. To obtain the performance of MetaMap, each instance of a target word in the NLM-WSD data set will be run through MetaMap. The target word in each instance will be assigned a UMLS concept. We will determine the accuracy of the system by comparing concepts tagged by MetaMap to the manually assigned concepts in the data set. To obtain the performance of our knowledge-based WSD system each instance of a target word in the NLM-WSD data set will be run through our system. The target word in each instance will be assigned a UMLS concept. We will determine the accuracy of the system by comparing concepts tagged by our system to the manually assigned concepts in the data set. This will be done twice, once using the sense inventory from the NLM-WSD data set and a second time using the sense inventory from the MetaMap *Candidate Retrieval* module. The accuracy of each system will be compared.

The second part is to evaluate our system against supervised WSD systems using the NLM-WSD data set. Possible systems include our supervised WSD system discussed in the preliminary work (see Section 4), and the supervised WSD systems introduced by (Leroy and Rindfleisch, 2004) and (Joshi, Pedersen, and Maclin, 2005).

The third part is to evaluate our system against knowledge-based WSD systems using the NLM-WSD data set. Possible systems include the system introduced by (Humphrey et al., 2006).

5.4 Possible Ways Forward

There are a number of ways forward in our the proposed research. Here, we discuss some possible ways that have arisen from our preliminary work and the related work that would contribute to the thesis. First, we discuss what information from the external data sources, UMLS and biomedical corpora, could be used and second, we discuss how that information can be used in the disambiguation process.

5.4.1 Knowledge Source Information

In our preliminary work, we evaluated using the surrounding UMLS concepts of the target word (Concept) and Journal Descriptors as a feature in our supervised WSD system. From this research and our discussion of related work, five questions arose that may be useful in our knowledge-based WSD system.

(Joshi, Pedersen, and Maclin, 2005) use unigrams as a feature in their supervised WSD system. Unigrams take into account how often a surrounding word is seen with the target word. We compared (Joshi, Pedersen, and Maclin, 2005) supervised WSD system to our supervised WSD system that used the UMLS concept of the words surrounding the target word (Concept) as a feature. We showed that Concept significantly improves on the overall baseline results by approximately 12% but does not improve upon the overall results reported by (Joshi, Pedersen, and Maclin, 2005). We hypothesized that it was possible that the farther abstracted away a feature gets from the individual word itself, the lower the accuracy obtained. We tested our hypothesis using the “bag-of-words” feature but found that Concept performed better than “bag-of-words” by 2%. Based on these results, we conclude that the frequency information used in the unigram feature may be important to the disambiguation process. This led us to our first question: Would using only the UMLS Concepts of the surrounding words that co-occur with the target word with a high frequency improve the results?

Words that co-occur together with a high frequency tend to be associated in some way. Statistical measures of association can be performed to determine the likelihood of two words occurring together. This led us to our second question: Would using only the UMLS Concepts of the surrounding words that are highly associated with the target word improve the results?

Words that co-occur together with a high frequency also tend to be related in some way. Similarity and relatedness measures can be performed to determine the relatedness of two words. (Caviedes and Cimino, 2004) introduce a measure to determine the similarity between concepts in the UMLS (see Section 3.3). This led us to our third question: Would using only the UMLS Concepts of the surrounding words that are related to the target word improve the results?

Semantic types were shown by (Leroy and Rindflesch, 2004) and (Humphrey et al., 2006) to perform well on the WSD task. Like Concepts, semantic types is another feature that may benefit from incorporating frequency, measures of association and/or similarity measures. This led us to our fourth question: Would

using only the semantic types of the surrounding words that are either highly related, highly associated or frequently co-occur with the target word improve the results?

A Journal Descriptor is a manually assigned tag to the journals in the National Library of Medicines' List of Serials Indexed for Online Users. The journals in the Serials include a majority of journals indexed in MEDLINE which is where the NLM-WSD data originates. In our preliminary work, we introduced Journal Descriptors as a potential feature for a supervised WSD system. We showed that using the Journal Descriptor as features may have potential but we can not say how much due to the limited number of journals in our data set that have a Journal Descriptors. Journal Descriptors are a coarse-grained feature. We felt that if they performed reasonably well in a supervised WSD system that they could be used in a coarse-grained filter to narrow down the number of senses in a knowledge-based WSD algorithm. This lead us to our fifth question: If we can get the NLM-WSD data set tagged with Journal Descriptors, how well would they perform in a coarse-grained filter?

5.4.2 Knowledge-based WSD Algorithm

The vector-based knowledge-based WSD systems (see Section 3.3.4 determines the sense of a word by taking the cosine between a target word vector and the possible sense vectors. The closest sense of the vector to the target vector (i.e. the sense vector with the smallest angle) is assigned to the target word. We believe that this approach may be a first step forward toward testing our above questions on a knowledge-based system. We would like though to investigate other ways of determining the distance between vectors that contain information about the words from the knowledge source or the information itself.

5.5 Proposed Contributions

Currently, there does not exist a WSD system that is both scalable and achieves high accuracy. If our proposed research is successful, a scalable WSD system that achieves reasonable accuracy would be beneficial a number of users. For example: the blind, global companies, Google users, and biomedical researchers.

A text-to-speech system for the blind that converts books from written text to speech would benefit from a knowledge-based WSD system. Text-to-speech is the task of producing the speech equivalent of written text. The appropriate sense of a word is needed pronounce some words properly. For example, the word "lead",

which is pronounced as [leed] to mean to take somebody somewhere and [led] to mean a toxic malleable metallic element. Such a system would allow a blind individual to listen to a newly released novel without waiting for it to become for it become a “book-on-tape”.

A machine translation system for global companies that converts technical documents from one language to another would benefit from a knowledge-based WSD system. Machine translation is the task of translating a text from one language into another such as German to English. The appropriate sense of a word is needed to translate it properly. For example, the German word “sicherheit” translates to “confidence” or “security”. Such a system would allow global companies to translate their product manuals from the standard, English, French and German to any language their customer speaks.

An information retrieval system such as Google would benefit from a knowledge-based WSD system. Information retrieval is the task of indexing, searching, and recalling data. The appropriate sense of a word is needed in order to query as well as return relevant documents to the user. For example, a set of documents all with the word “bat” should be indexed based on whether the document is talking about the “bat” that flies or the baseball instrument. When querying for documents about mammals that fly at night, documents about baseball would not be returned.

A concept-mapping system such as MetaMap would benefit from a knowledge-based WSD system. Concept mapping is the task of automatically linking a document to concepts (senses) in an ontology. The mapping is done by linking content words in the document to their appropriate concept in the ontology. In order to do this accurately, the appropriate concept needs to be identified. The concept mapping task is a very similar to the information retrieval task. A researcher could query articles that contain one set of concepts but do not contain another. For example, a researcher could query for articles about the protein “interferon- α ” specifying that the articles contain information about “new data on the expression levels in lymphatic tissue” but not “new methods to get more expression levels”.

References

- Agirre, E. and G. Rigau. 1996. Word sense disambiguation using Conceptual Density. *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 16–22.
- Agirre, Eneko and Philip Edmonds. 2006. *Word Sense Disambiguation Algorithms and Applications*. Springer.
- Altintas, E., E. Karşigil, and V. Coskun. 2005. A new semantic similarity measure evaluated in word sense disambiguation. In *15th NODALIDA conference*, Joensuu, January.
- Aronson, A R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21.
- Aronson, A.R., O. Bodenreider, H.F. Chang, S.M. Humphrey, J.G. Mork, S.J. Nelson, T.C. Rindfleisch, and W.J. Wilbur. 2000. The NLM Indexing Initiative. *Proc AMIA Symp*, 20:17–21.
- Banerjee, S. 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.
- Bhattacharya, I., L. Getoor, and Y. Bengio. 2004. Unsupervised sense disambiguation using bilingual probabilistic models. *Meeting of the Association for Computational Linguistics*.
- Briscoe, T. and J. Carroll. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. *Arxiv preprint cmp-lg/9510005*.
- Bruce, R. and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. *Proceedings of the 32nd conference on Association for Computational Linguistics*, pages 139–146.
- Budanitsky, A. and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources*, 2.
- Cabezas, Clara, Indrajit Bhattacharya, and Philip Resnik. 2004. The university of maryland senseval-3 system descriptions, July.
- Caviedes, J.E. and J.J. Cimino. 2004. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85.
- Cowie, J., J. Guthrie, and L. Guthrie. 1992. Lexical disambiguation using simulated annealing. *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 1:359–365.
- Dempster, AP, NM Laird, and DB Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

- Diab, M. and P. Resnik. 2001. An unsupervised method for word sense tagging using parallel corpora. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262.
- Diab, Mona. 2000. An unsupervised method for multilingual word sense tagging using parallel corpora: a preliminary investigation. In *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.
- Fellbaum, Christiane. 1998. *WordNet – An electronic lexical database*. Cambridge, Massachusetts and London, England: MIT Press.
- Gale, W., K. Church, and D. Yarowsky. 1992a. One sense per discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237.
- Gale, W.A., K.W. Church, and D. Yarowsky. 1992b. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439.
- Gale, William, Kenneth Ward Church, and David Yarowsky. 1992c. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 249–256, Morristown, NJ, USA. Association for Computational Linguistics.
- Hirst, G. and D. St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An Electronic Lexical Database*, pages 305–332.
- Humphrey, SM, TC Rindflesch, and AR Aronson. 2000. Automatic indexing by discipline and high-level categories: Methodology and potential applications. *Proceedings of the 11th ASIST SIG/CR Classification Research Workshop*, pages 103–116.
- Humphrey, Susanne M., 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.
- Ide, N. and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):2–40.
- J. Jiang, D. Conrath. 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.
- Joshi, M., 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 *Second Indian International Conference on Artificial Intelligence*, Pune, India, December.

- Kilgarriff, A. and J. Rosenzweig. 2000. English SENSEVAL: Report and Results. *LREC, Athens*.
- Klapaftis, I. and S. Manandhar. 2005. Google and wordnet based word sense disambiguation. *Proceedings of the 22nd International Conference on Machine Learning Workshop on Learning and Extending Ontologies by using Machine Learning Methods*.
- Leacock, C. and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*, pages 265–283.
- Leacock, C., G.A. Miller, and M. Chodorow. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.
- Leacock, C., G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. *Proceedings of the workshop on Human Language Technology*, pages 260–265.
- Lee, Y.K. and H.T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 41–48.
- Lee, Y.K., H.T. Ng, and T.K. Chia. 2004. Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain. Association for Computational Linguistics*.
- Leroy, G. and Thomas C. Rindfleisch. 2004. Using Symbolic Knowledge in the UMLS to Disambiguate Words in Small Datasets with a Naive Bayes Classifier. *MEDINFO*.
- Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.
- Li, H. and N. Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.
- Lin, D. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 64–71.
- Liu, H., V. Teller, and C. Friedman. 2004. A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation.
- Manning, C. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

- McCarthy, D. and J. Carroll. 2003. Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences. *Computational Linguistics*, 29(4):639–654.
- McRoy, S.W. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.
- Mihalcea, R. 2002. Bootstrapping large sense tagged corpora. *Proceedings of the Third International Conference on Language Resources and Evaluation LREC 2002*.
- Mihalcea, R. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. *Proc. of EMNLP2005*.
- Mihalcea, R. and D. Moldovan. 2000. An iterative approach to word sense disambiguation. *Proceedings of FLAIRS*, pages 219–223.
- Mihalcea, Rada and Dan I. Moldovan. 1999. A method for word sense disambiguation of unrestricted text. pages 152–158.
- Mohammad, S. and G. Hirst. 2006. Determining Word Sense Dominance Using a Thesaurus. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Mohammad, S. and T. Pedersen. 2004. Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation. *Proceedings of the Eighth Conference on Natural Language Learning at HLT-NAACL*.
- Molina, A., F. Pla, E. Segarra, and L. Moreno. 2002. Word Sense Disambiguation using Statistical Models and WordNet. *Proceedings of 3rd International Conference on Language Resources and Evaluation, LREC2002, Las Palmas de Gran Canaria, Spain*.
- Mooney, R.J. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82–91.
- Navigli, R. and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086.
- Ng, H.T. 1997. Exemplar-based word sense disambiguation: Some recent improvements. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 208–213.
- Ng, H.T. and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An

- exemplar-based approach. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 47.
- Patwardhan, S. 2003. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. *Master's thesis, Univ. of Minnesota, Duluth*.
- Pedersen, T. 2000. A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, 1.
- Pedersen, T., S. Banerjee, and S. Patwardhan. 2005. Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. *Supercomputing institute research report umsi*, 25.
- Pedersen, T. and R. Bruce. 1997. Distinguishing word senses in untagged text. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 2:197–20.
- Pedersen, T. and R. Bruce. 1998. Knowledge lean word-sense disambiguation. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 800–805.
- Pedersen, T., S.V. Pakhomov, S. Patwardhan, and C. Chute. 2006. Measures of semantic similarity and relatedness in the biomedical domain. *Biomedical Informatics, Elsevier*.
- Pedersen, Ted. 2004. The duluth lexical sample systems in senseval-3. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 203–208, Barcelona, Spain, July. Association for Computational Linguistics.
- Purandare, A. 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.
- Rada, R., 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.
- Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1:448–453.
- Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. MEANING: a roadmap to knowledge technologies. *International Conference On Computational Linguistics*, pages 1–7.
- Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
- Snyder, Benjamin and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil

- Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.
- Stevenson, M. and Y. Wilks. 2001. The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics*, 27(3):321–349.
- Sussna, Michael. 1993. Word sense disambiguation for free-text indexing using a massive semantic network.
- Widdows, D., S. Peters, S. Cederberg, C.K. Chan, D. Steffen, and P. Buitelaar. 2003. Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. *Natural Language Processing in Biomedicine ACL 2003 Workshop*, pages 9–16.
- Witten, I.H. and E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.
- Wu, Z. and M. Palmer. 1994. Verbs semantics and lexical selection. *Proceedings of the 32nd conference on Association for Computational Linguistics*, pages 133–138.
- Yarowsky, D. and R. Florian. 2003. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(04):293–310.
- Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora. *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 454–460.
- Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd conference on Association for Computational Linguistics*, pages 189–196.