Supervised and Knowledge-based Methods for Disambiguating Terms in Biomedical Text using the UMLS and MetaMap

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Bridget Thomson McInnes

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor Of Philosophy

September, 2009

Acknowledgements

This dissertation would not be what it is without the help from a number of individuals. I want to thank:

- Ted Pedersen, who can't quite seem to get rid of me. Thank you for your guidance, support and continually asking the question 'why'.
- John Carlis for taking me on as an advisee even though I am not fond of his candy and much prefer chocolate.
- Serguei Pakhomov and Arindam Banerjee for being on my thesis committee and providing valuable feedback.
- Lan Aronson, Jim Mork, Will Rogers and François Lang, from the National Library of Medicine, for their help in understanding the tools provided by the library and the wonderful time I had while working there.
- Kin Wah, Jan Willis and Olivier Bodenreider, also from the National Library of Medicine, for their help in understanding the UMLS. The thing is a bear and I would never have been able to wrap my head around it without their willingness to answer my seemingly endless set of questions.
- Shana Watters, Sara Drenner and Getiria Onsongo, my comrades in crime.
- My parents, Kathleen and Gordon, for their continually encouragement.
- Chris Buck for his patience and support through this entire process.

Dedication

To Chris Buck. Thank you.

Supervised and Knowledge-based Methods for Disambiguating Terms in Biomedical Text using the UMLS and MetaMap

by Bridget Thomson McInnes
Under the supervision of Dr. Ted Pedersen and Dr. John Carlis

ABSTRACT

Word Sense Disambiguation is the task of automatically identifying the appropriate sense (or concept) of an ambiguous word, for example, the term *cold* could refer to the temperature or a virus depending on the context in which it is used. Not being able to identify the intended concept of an ambiguous word negatively impacts the accuracy of biomedical applications such as medical coding and indexing which are becoming essential in the biomedical and clinical world with the push towards electronic medical records and the growing amount of information that is available to biomedical researchers and clinicians. This dissertation focuses on disambiguating ambiguous words in biomedical text.

This dissertation presents two methods, K-CUI and A-CUI, that can disambiguate ambiguous terms in any biomedical text using information from the Unified Medical Language System (UMLS). K-CUI explores the use of Concept Unique Identifiers (CUIs) as assigned by MetaMap, as features for a supervised learning method for word sense disambiguation. It also investigates four techniques to reduce the noise in the feature set by restricting which CUIs to include. The first technique is windowing, whose results show that in biomedical text indicative CUIs are highly localized. The second is a frequency cutoff, whose results show that when a dataset contains a high majority concept, the features that only occur a few times are essential in disambiguating the minority concepts. The third is a MetaMap Indexing cutoff, whose results show that word concepts are correlated with the topical information describing an instance. The fourth is a semantic similarity cutoff, whose results show in biomedical text, indicative features have a high semantic similarity with at least one of the possible concepts of the ambiguous word.

A-CUI is a knowledge-based method that uses information from the UMLS and MetaMap mapped text to represent the context of the possible concepts of an ambiguous word. It investigates three types of contextual representations. The first uses the concept's definition in the UMLS, whose results show that the context used with the words the definition can be used to represent its context of the concept. The second uses the preferred and associated terms from the UMLS, whose results show that the terms themselves do not provide enough contextual information to disambiguate between the possible concepts of a target word. The third uses the words surrounding the concept in MetaMap mapped text, whose results show that the information provided by MetaMap is distinct enough to distinguish between the possible concepts for disambiguation purposes.

K-CUI and A-CUI are evaluated using the NLM-WSD dataset which consists of Medline abstracts. Previous work in this area have also evaluated their methods using the same dataset and in some cases tailored their methods to work only on Medline abstracts. Identifying the correct concept of an ambiguous term in Medline abstracts is a significant problem but the advantage of K-CUI and A-CUI though is that they are portable systems that can disambiguate terms in any biomedical text, unlike previous methods that are limited to only Medline abstracts.

There has also been previous work that determines the correct concept of a target word by first identifying the target words semantic type which is a broad categorization of a concept. After the semantic type of the ambiguous words is identified, then the correct concept is identified based on its semantic type. The assumption is that each possible concept of a target word has a unique semantic type. If the possible concepts have the same semantic type this method cannot distinguish between them; A-CUI and K-CUI do not have the limitation. Also, identifying the semantic type of a target word is a simpler problem than identifying the concept because semantic types are a coarser grained categorization than CUIs which makes them easier to assign.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Historically, communication between humans has been face to face. Any ambiguity that arose during the transfer of information was solved as the conversation continued. Technology has changed the way that people communicate and information is now stored for retrieval at a later date. In order to retrieve this information, the computer must be able to resolve any ambiguity that arises without the aid of the creator of the document.

Consider the following phrase:

zinc gluconate lozenges are for treating the common cold in children

As humans, each word in this sentence has a clear meaning. It is obvious that the word *cold* is referring to an upper respiratory infection, although there are other possible meanings such as an absence of heat, sensation produced by low temperatures, feeling or showing no enthusiasm, or the state of unconsciousness. This list is far from conclusive and yet the correct meaning of *cold* was identified on the fly without much thought or difficulty.

For the computer, though, this distinction is not so obvious and it finds it difficult to disambiguate between these different meanings. Word Sense Disambiguation (WSD) is the task of automatically identifying the correct meaning of a word that has multiple meanings. In WSD, these meanings are referred to as *senses*, or *concepts*, which are obtained from a *sense-inventory*. The ambiguous word is referred to as the *target word* and the context in which the target word is used is called an *instance*. In the example of above, the instance consists of a phrase, but it could just as easily be a sentence,

phrase or paragraph.

Research in WSD began in the 1940's with the publication of an influential memorandum by [Weaver, 1949]. In this document, he identified WSD as one of the main problems of translating one language into another (Machine Translation), since the different concepts of a word in one language may be translated to entirely different words in another. For example, in French, the word *grille* can be translated to *railings*, *gate*, *bar*, *scale* or *schedule* depending on the context in which it is used. The problem of word ambiguity was identified as a significant problem early in computer history.

WSD is important to many tasks other than Machine Translation. Examples include text-to-speech, information retrieval and concept mapping. Text-to-speech is the task of producing the speech equivalent of written text. Examples of a text-to-speech system are automatic announcement systems such as those for weather, airport arrivals/departures, or movie showings. The appropriate concept of a word is needed to pronounce some words properly. For example, the word *bass*, pronounced [beys], to mean a low pitched singing voice ,or [bæs], to mean the fish.

Information retrieval is the task of indexing, searching, and recalling data. Documents need to be properly indexed based on the concept of the words in the documents rather than the word itself in order for the appropriate documents to be returned.

Concept mapping is the task of automatically linking documents to concepts in a lexical database which is done by *mapping* content words in documents to their appropriate concept in the database. In order to do this accurately, the appropriate concept must be identified. One such system is MetaMap which maps terms in biomedical text to concepts in the Unified Medical Language System. [Aronson, 2001] notes that a WSD component would greatly improve the accuracy of their system called MetaMap.

Although, research in this area has now been going on for at least 70 years, it is still considered a difficult problem that has not been solved satisfactorily. [Navigli, 2009] notes that the difficulty of creating an accurate WSD system that could be used by other natural language processing systems still exists. They state:

> "The identification of the specific meaning that a word assumes in context is only *apparently* simple."

This dissertation focuses on the disambiguating terms in biomedical text which is a relatively new area. The text consists of journal articles and abstracts about topics

including but not limited to life science, anatomy, biology, and biochemistry. There are additional types of biomedical text such as clinical reports but currently most of the research conducted in this area focuses on biomedical journal articles due to the availability of evaluation datasets. The methods proposed in this dissertation are evaluated on biomedical journal articles but can be used to disambiguate terms in other types of biomedical text.

Few researchers have done work specifically using biomedical information to disambiguate words in biomedical journal articles. Most use previous methods that have been used to disambiguate words in general English. These methods obtain a high disambiguation accuracy, perhaps because general English methods use the context surrounding the target word to determine its correct concept. These methods calculate the probability of a word being in the same context as the target word to determine the appropriate concept. The probability is obtained by counting the number of times a word was seen with the target word in some text, referred to as the training data. Applying this method to biomedical text requires only a reference text in the biomedical domain to determine the probability of biomedical terms occurring with other biomedical terms.

Recently there has been some work looking at biomedical information from the Unified Medical Language System (UMLS) to determine if it can be leveraged to help in the disambiguation process. The UMLS is a framework that integrates concepts from biomedical and clinical sources into a single database containing syntactic and semantic information about those concepts. It is produced and maintained by the National Library of Medicine.

This dissertation seeks to use information from the UMLS Metathesaurus which is a lexicon containing biomedical and clinical concepts. The Metathesaurus contains over 1.5 million concepts called Concept Unique Identifiers (CUIs) which come from various biomedical and clinical sources integrated through a combination of electronic and human effort into a single source. As of version 2008AB, there exist over 100 sources that have been integrated.

CUIs provide unambiguous term level information that may not be captured by the term itself. CUIs in the UMLS are mapped to terms rather than individual words. For example, the term $falling\ down$ is mapped to the concept Falls [C0085639] creating a single concept for the term rather than two concepts for each of the words in the term.

A CUI is expressed by having specific attributes that define it, such as its definition, and the terms used to refer to it. For example, $fall$, referring to the concept of not remaining upright, is CUI Falls [C0085639] in the UMLS version 2008AB which have the following associated terms:

- fall
- falls
- falls down
- falling
- falling down

This type of information has not been previously used in WSD systems. This dissertation proposes two novel methods to disambiguate words in biomedical text, K-CUI and A-CUI, which uses CUI information extracted from the UMLS and MetaMap mapped text to determine the appropriate concept of an ambiguous word.

K-CUI is a supervised method that uses information from a dataset where each instance in the dataset is manually annotated with its correct concept. The novelty of K-CUI is its use of CUIs assigned by MetaMap of the words surrounding the target word. These CUIs are to determine the correct concept target word.

A-CUI is a knowledge-based method which uses CUI information from the UMLS and biomedical text that has been automatically mapped to CUIs by MetaMap to determine the appropriate concept. The novelty of A-CUI is that it creates a contextual description of the concept using CUI information and then compares this context with the instance containing the target word.

The K in K-CUI stands for 'kid' because kids require supervision and the A in A-CUI stands for 'adult' since they do not.

The basic contributions of this dissertation are:

- Using the CUIs of the words surround the target word as feature of a supervised learning algorithm.
- Using information from the UMLS and MetaMap mapped text to provide a contextual description for a concept in a novel, knowledge-based WSD method.

A more detailed list of the contributions are discussed in Chapter 9.

The remainder of the dissertation is organized as follows. Chapter 2 discusses some fundamental concepts that are required to understand WSD and how it is pursued in this dissertation.

Chapter 3 discusses the proposed K-CUI system. This chapter first describes the system and discusses the implementation details. Second, an evaluation is conducted of the system using the NLM-WSD dataset which is a biomedical text in which the ambiguous words in the text have been manually annotated with their appropriate concept. Lastly, it discusses of the results of the evaluation. Chapter 5 contains a similar discussion of the proposed A-CUI system. First, it describes the system and discusses the implementation details, second, an evaluation is conducted using the NLM-WSD dataset, and lastly, it discusses of the results of the evaluation.

Chapter 7 discusses the related work that has directly contributed to this dissertation. WSD has a long history and there are a number of different methods that have been proposed. The previous work included in this section are those that have a direct relation to the work that has been conducted in the biomedical domain, this includes methods which were originally created to disambiguate general English terms and were later used to disambiguate biomedical terms as well as those system that were created specifically to disambiguate biomedical terms.

The overall results of K-CUI and A-CUI indicate other avenues of research. Chapter 8 discusses potential future work in biomedical WSD. Lastly, Chapter 9 discusses the specific contributions and the overall conclusions of this dissertation.

Chapter 2

Background

In order to understand WSD as a problem as pursued in this dissertation there are a few fundamental concepts that must be first presented.

First, the methods in this dissertation used numeric vectors, called feature vectors, to represent the context in which a target word is used. There are two types of feature vectors, *first-order* and *second-order*, vectors which fall into three different categories:

- concept vectors
- test vectors
- training vectors

Second, this dissertation discusses three methods used for WSD:

- supervised WSD methods
- clustering WSD methods
- knowledge-based WSD methods

Supervised methods use manually annotated training data containing instances of a target word to learn the context in which target words are used. Then, when a new instance of the target word is seen, the correct concept is determined based on its context compared to the context of the previously seen instances. Manually annotated training data contains instances of a target word, called training instances, that have been manually assigned their correct concept by humans. Clustering methods use unannotated training data containing instances of the target word. Unannotated training data

contains training instances whose concepts are not known. These instances are grouped together to form clusters where each cluster contains only those instances that have the same concept. Knowledge-based methods use information extracted from curated and structured data called a knowledge source. These methods rely on information from the knowledge source about a concept such as its definition or synonym rather than training instances in manually annotated or unannotated training data.

Third, this dissertation refers to two knowledge sources that have been used in WSD: WordNet and the Unified Medical Language System (UMLS). These knowledge sources are domain specific and contain curated information about their domain. WordNet is a lexical database that contains concepts from the general English domain and the UMLS is a lexical database that contains concepts from the biomedical domain.

Fourth, this dissertation uses and refers to several training datasets. Some of these datasets are manually annotated specifically for WSD and others are unannotated. The manually annotated datasets contain instances of a target word that have been manually assigned a concept from a specific sense-inventory and were created for the purpose of evaluation and training data for supervised WSD methods.

Fifth, this dissertation incorporates three software packages into its experimental framework: MetaMap, the WEKA data mining package, and SenseClusters. MetaMap is used to extract biomedical information about terms from the UMLS. The WEKA data mining package contains supervised learning algorithms that can be used in supervised WSD methods. SenseClusters is a clustering WSD system that provides software programs to create first and second-order feature vectors.

The remainder of this chapter discusses these fundamental concepts in more detail.

2.1 Feature Vectors

WSD methods use feature vectors to represent the context in which a target word is used. The context of the target word is what is used by WSD methods in order to disambiguate between possible concepts it represents.

[Miller and Charles, 1991] show that the similarity between words can be determined based on the similarity between their contexts. Words that have a similar meaning can be substituted in a sentence without changing the meaning of the sentence or making

it nonsensical. Conversely, substituting words that are not synonymous leads to nonsensical sentences that would never be used. The assumption is that words with similar meaning will have similar contexts and words that do no have a similar meaning will not. This assumption can be applied to WSD. Consider the following example of the target word *cold*:

He caught a *cold* that winter

The word *cold* has multiple meanings under which two of them are: i) a low temperature and ii) the virus. Substituting low temperature for the virus results in a grammatical but nonsensical sentence.

Feature vectors represent the context of a target word as an n-dimensional vector of numerical features, called a feature set, where a feature provides a distinction between concepts for classification. Features are extracted from training data or a knowledge-source. For example, consider the following training instances containing the target word *cold*:

- The groups susceptibility to a cold appeared to be positively associated with the risk.
- He used a combination of the UW solution both for initial flush and the cold immersion.

where cold refers to the $Common\ Cold$ in the first instance and $Cold\ Temperature$ in the second. The features are extracted from these instances to create feature vectors. One example of a feature set is called $bag\text{-}of\text{-}words$ which uses the content words in the training data as features. The content words are identified using a stoplist which is a list that contains non-content words such as determiners and prepositions. A word is included in the feature set if it is not listed in the stoplist. Using the above training instances, the feature set would contain the following:

- used
- combination
- UW
- solution
- initial

- flush
- immersion
- groups
- susceptibility
- appeared
- positively
- associated
- risk

The elements in the feature vectors are numerical indicators as to the existence or non-existence of the feature in the instances. These elements could refer to the number of times the feature occurs in the training data, or the association between the feature and the target word. The elements in this example are a one or zero indicating whether or not a feature exists in the same instance as the target word. The feature vector for each of the instances in the training data are shown in Figure 2.1 and 2.2.

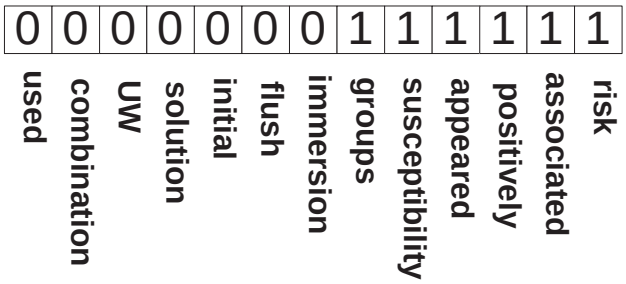

Figure 2.1: Feature Vector for Instance 1

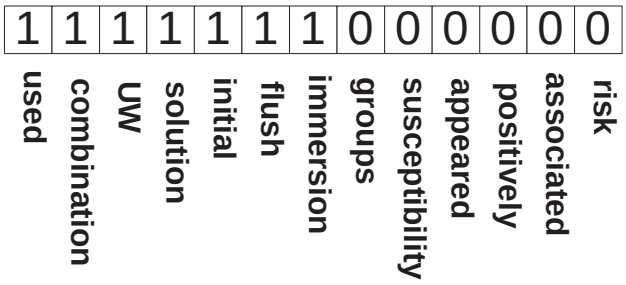

Figure 2.2: Feature Vector for Instance 2

There exist other types of features that can be included in the feature set, one such example is the part-of-speech of the target word which can be used to disambiguate between possible concepts of a target word. Consider the following instances of the target word *cold*:

- He caught a *cold*
- He caught a *cold* fish

In the first instance, *cold* is a noun and refers to the concept "Common Cold" but in the second instance, *cold* is an adjective and refers to the concept "Cold Temperature". Chapter 7 discusses the other features that have previously been used for WSD in more detail.

This dissertation classifies feature vectors into three categories: training vectors, test vectors and concept vectors. The vectors in these categories can be one of two types: first-order vectors or second-order vectors. Hence, six different vectors can be created:

- first-order training vector
- first-order test vector
- first-order concept vector
- second-order training vector
- second-order test vector
- second-order concept vector

The following two subsections describe the three categories of feature vectors and then the two different ways to create them.

Categories of Feature Vectors

The feature vectors described above are referred to as training vectors. These are feature vectors of training instances in manually annotated or unannotated training data. The concept of a training vector created from manually annotated training data is known whereas a concept of a training vector created from unannnotated training data is not known.

There are two other categories of feature vectors used in this dissertation: test vectors and concept vectors. A *test vector* is a feature vector of an instance in the test

data called a test instance. A test vector contains features from the training data and its elements are numerical indicators of the existence or non-existence of the feature in a test instance. For example, consider the following test instance:

The control group consisted of a cold flush with heparinzed solution

And the following training instances from the previous example:

- The groups susceptibility to a cold appeared to be positively associated with the risk.
- He used a combination of the UW solution both for initial flush and the cold immersion.

Figure 2.3 shows the test vector for this instance using the *bag-of-words* feature set. The elements of the vector are either a one or a zero indicating whether or not the feature occurs in the test instance. It is important to realize that the features for both the training and test vectors are exactly the same and come only from the training data. The words such as *wind* and *rain* in the test instance do not exist in the training data therefore are not included in the feature set.

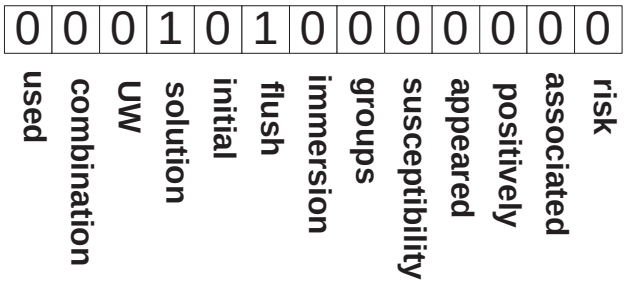

Figure 2.3: Test Vector

The concept vector is the representation of a possible concept of the target word originating from [Schütze, 1992]. A concept vector is a generalization of the context in which the concept can be used. The methods discussed in this dissertation create the concept vectors using two different techniques. The first is by calculating the centroid of a set of training vectors that have been labeled (either automatically or manually) with the concept. This creates a concept vector whose non-zero elements are features that occur with that concept in the training data.

The second is by obtaining a context describing the concept and using this context as its instance. The concept vector consists of features from the training data and its elements are numerical indicators of the existence or non-existence of the feature in its context. An example of such a context is the concept's definition. Consider the two possible concepts for the target word *cold* and their corresponding definitions:

- *Common Cold*: a contagious, viral infection of the respiratory system; there known cure, but it is not fatal for low risk groups.
- *Cold Temperature*: an absence of warmth or heat or a temperature notably below an accustomed norm; a measure of the average kinetic energy of the particles in a sample solution.

And the following training instances from the previous example:

- The groups susceptibility to a cold appeared to be positively associated with the risk.
- He used a combination of the UW solution both for initial flush and the cold immersion.

Figures 2.4 and 2.5 show the concept vectors for each of the concepts. The feature set consists of the words in the training instances and the elements are either a one or a zero indicating whether or not the feature occurs in the concepts definition.

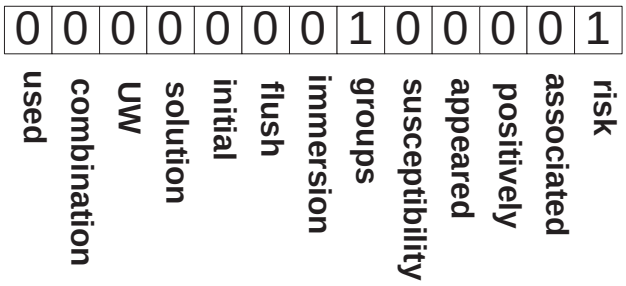

Figure 2.4: Concept Vector for Common Cold

Figure 2.5: Concept Vector for Cold Temperature

First-order and Second-order Vectors

Two different types of vectors are used in this dissertation: first-order and second-order. The feature vectors shown in the previous examples are first-order vectors. In first-order vectors the features come from the training data and the elements are numerical indicators as to the existence or absence of that feature in its associated context.

The size of the vector depends on the number of features that exist in the training data. The vectors in the previous example are very small but can becomes larger given more training instances. The disadvantage of these vectors is that they consist predominately of zeros. Second-order vectors attempt to alleviate the sparseness. The elements in these vectors are a numerical indicator as to whether or not the feature was seen with a word in the associated context not just with the target words.

Second-order vectors are created by first creating a first-order vector for each content words in the instance. The features in the first order vector come from the training data and the elements are numeric indicators of whether or not the content word occurs with the feature in the training data. Then these first-order vectors are averaged together to create a second-order vector.

Consider the following test instance containing the target word cold.

The control group consisted of a cold flush with heparinzed solution.

A first-order vector is created for each of the content words in the instance: *control*, *group*, $flush$, *heparinzed*, and *solution*. The features in these vectors come from the training data and the elements are a one or a zero indicating whether or not the content word occurs with the feature in the training data. A second-order test vector is then

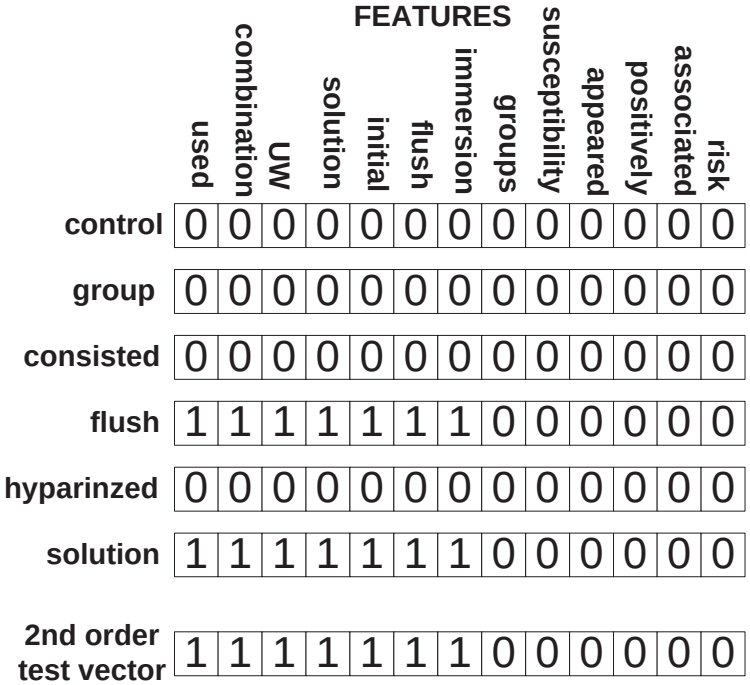

Figure 2.6: Second-Order Test Vector

created by averaging all of the first-order vectors. Figures 2.6 shows the second-order test vector for this test instance.

Now consider the concept "Common Cold" and its following definition:

A contagious, viral infection of the respiratory system; there known cure, but it is not fatal for low risk groups.

A first-order vector is created for each of the content words in the definition: *contagious*, *viral*, $infection$, *respiratory system*, *cure*, $fatal$, *low*, *risk*, and *groups*. The features in these vectors come from the training data and the elements are a zero or a one indicating whether or not the content word occurs with the feature in the training

data. A second-order concept vector is then created by averaging all of the first-order vectors. Figures 2.7 shows the second-order concept vector for the concept "Common Cold". This dissertation creates the first and second-order vectors using the vector programs in SenseClusters which is discussed in more detail in Section 2.4.3.

2.2 Word Sense Disambiguation Methods

This section presents a general description of three WSD methods: supervised, clustering and knowledge-based methods.

2.2.1 Supervised WSD Methods

Supervised WSD methods rely on the use of manually annotated training data. The instances in the training data are manually annotated with their appropriate concepts from a sense-inventory. A supervised learning algorithm *learns* to recognize the context surrounding these concepts, creating a model which is used to automatically assign concepts to instances containing the target word in the test data.

Supervised learning methods in general obtain a very high disambiguation accuracy, outperforming other WSD methods. The disadvantage of these methods though is that they require manually annotated training data for each word that needs to be disambiguated. This is a labor intensive and time consuming process. There has been work in trying to automatically create the training data such as the method described by [Yarowsky, 1995] using general English and more currently by [Fan and Friedman, 2008] in the biomedical domain.

Figure 2.8 shows a general model of supervised WSD methods. In this method, an evaluation module takes manually annotated training data as input and splits the data into a training and test portion. The concept information is removed from the test portion and then both the datasets are sent to the vector creation program. The vector creation module extracts the features from the training data and creates a training vector for each instance in the manually annotated training data and a test vector for each instance in the test data. A supervised learning algorithm takes the training vectors as input and learns the context in which each of the possible concepts is used.

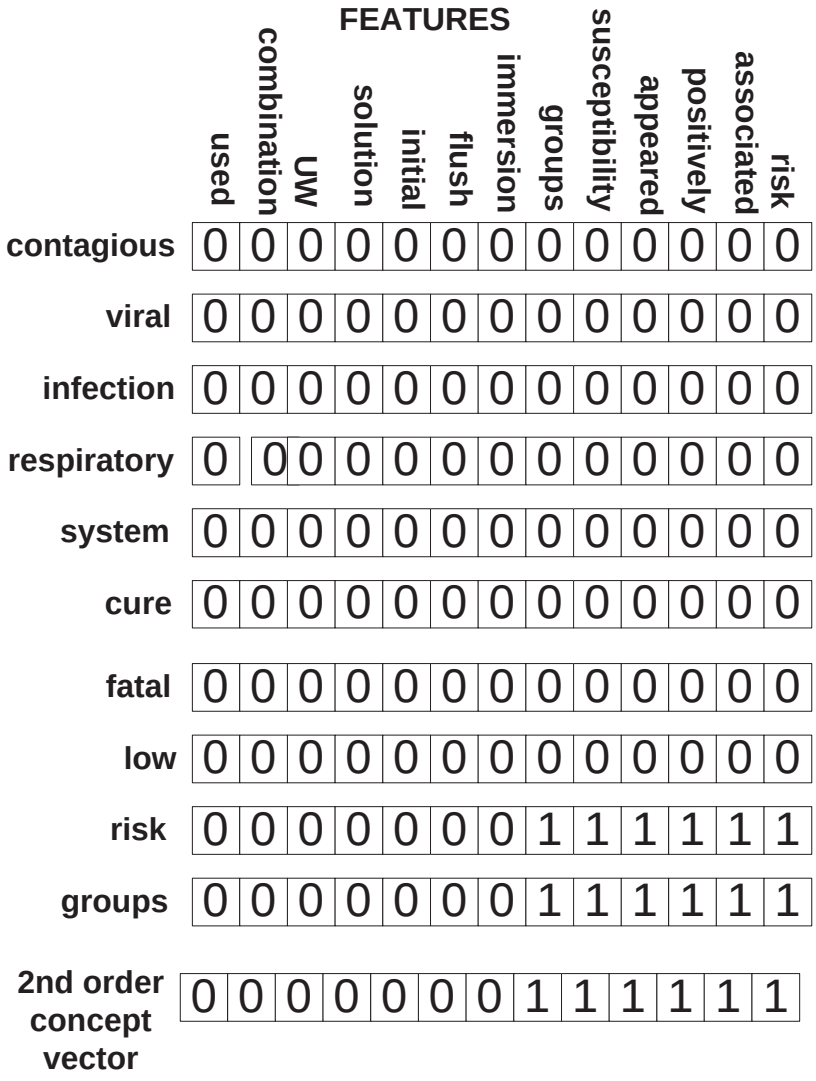

Figure 2.7: Second-Order Concept Vector

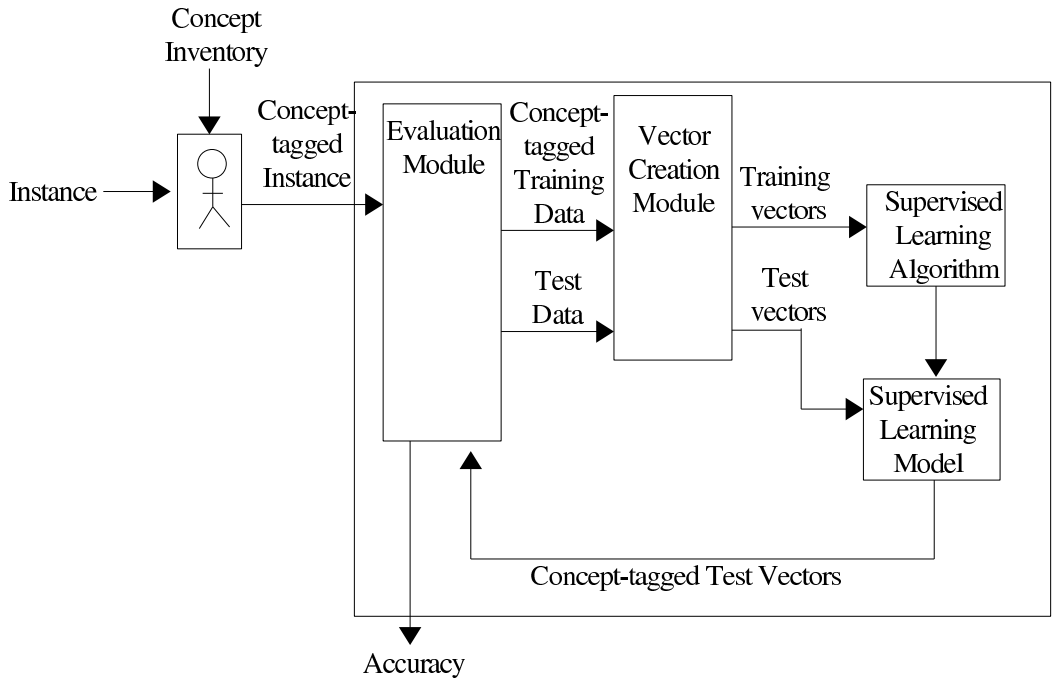

Figure 2.8: Supervised WSD Method

The algorithm creates a model which takes the test vectors as input and assigns each of the vectors their appropriate concept. The evaluation program takes these vectors as input and calculates the accuracy of the model.

There are a number of different supervised learning algorithms that have been used in supervised WSD methods. This section describes two supervised learning algorithms that have been used to disambiguate words in the biomedical domain:

- Support Vector Machines (SVMs)
- Naive Bayes classifier

Support Vector Machines (SVM)

SVMs identify the appropriate concept of a test instance based on where its vector lies in relation to the training vectors in some n-dimensional space. In this algorithm, a training vector is created for each instance in the manually annotated training data and mapped to an n-dimensional space. The assumption is that training vectors annotated

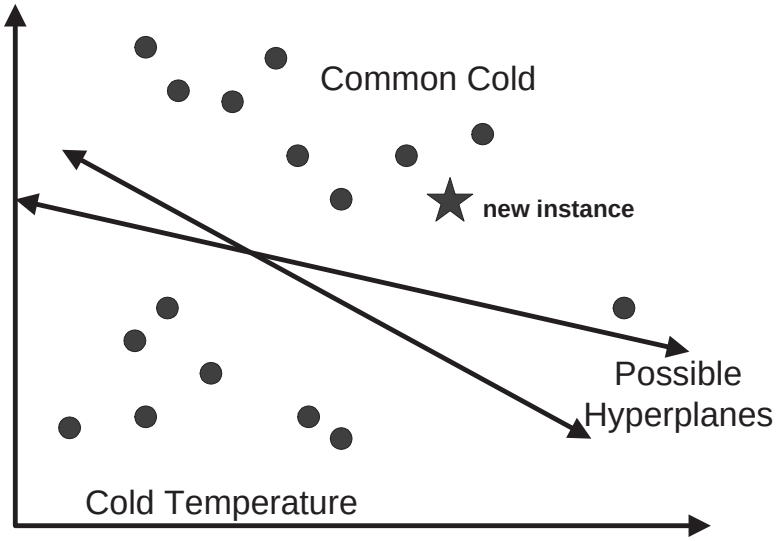

Figure 2.9: Support Vector Machine Possible Hyperplanes Example

with the same concept will be situated together. For example, consider the target word *cold* which has two possible concepts:

- *Common Cold*
- *Cold Temperature*

The training vectors are mapped onto an n-dimension space and the algorithm separates the vectors by creating a hyperplane such that the training vectors assigned the concept "Common Cold" are on one side of the hyperplane and the training vectors assigned the concept *Cold Temperature* are on the other. A hyperplane defines a k-dimensional subspace within an n-dimension space. For example, a line is a two-dimension hyperplane within an n-dimensional space.

A number of different possible hyperplanes that could be created to separate the training vectors as seen in Figure 2.9. The SVM creates a hyperplane such that the largest distance between both sets of vectors is maximized; this space between the hyperplane and the vectors is called the *margin*. To determine the margin, two parallel

hyperplanes, called *support vectors*. The goal is to create the support vectors as close to the training vectors as possible in order to obtain the largest margin and then create the hyperplane directly between them. A test vector is assigned a concept by mapping it onto the n-dimensional space and determining what side of the hyperplane it lies. An example of this is illustrated in Figure 2.10.

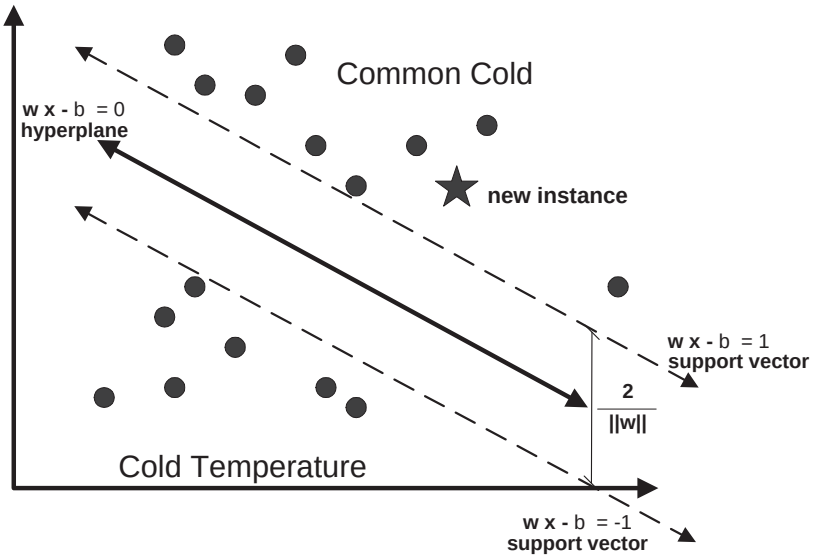

Figure 2.10: Support Vector Machine

The problem of determining the hyperplane is mathematically defined as follows:

$$\mathcal{D} = \{(\mathbf{x}_i, c_i)|\mathbf{x}_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^{n} \tag{2.1}$$

where the c_i is either 1 or -1. The support vectors are defined the set of points $\mathbf{x}$ satisfying the following equation:

$$\mathbf{w} \cdot \mathbf{x} - b = 1 \tag{2.2}$$

$$\mathbf{w} \cdot \mathbf{x} - b = -1. \tag{2.3}$$

The goal is to choose $\mathbf{w}$ and b to maximize the margin between the support vectors such that they are as far apart as possible while still separating the data. The hyperplane is mathematically defined as:

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \tag{2.4}$$

This equation draws the hyperplane directly between the support vectors so there is an equal distance between both.. The vector $\mathbf{w}$ is a normal to the hyperplane; meaning it runs perpendicular to it. The value $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. Figure 2.10 shows an example of the support vectors and hyperplane separating training vectors that have been assigned either the concept *Common Cold* or *Cold Temperature*.

For a linear classifier the two support vectors are selected such that there are no training vectors between them and the distance is minimized. The distance between the two support vectors is $\frac{2}{\|\mathbf{w}\|}$. Therefore the goal is to find the smallest $\|\mathbf{w}\|$ that satisfies the above Equations 2.2 and 2.3.

For a non-linear classifier, such as one that is classifying a set of data points that lie in and outside a circle, a "kernel trick" is used. The kernel trick is a method that uses a linear classifier to solve a non-linear problem by mapping vectors to a higher-dimensional space and then use the linear classification. The kernel trick depends on the dot product between two vectors. So whenever a dot product is used, it is replaced by a kernel function; this replacement is what is considered the trick.

The SVM described above is a binary classifier. This method is extended for multi-classification by looping through the possible concepts and classifying the instance as a possible concept or not. For example, if there exist three possible concepts, *Common Cold*, *Cold Temperature* and Cold Therapy, the SVM first classifies the instances as *Common Cold* or not-*Common Cold*, and then classifies all the instances of not *Common Cold* as *Cold Temperature* or *Cold Therapy*. This dissertation uses the linear SVM, SMO, from the WEKA data mining package which is described in more detail in Section 2.4.2.

Naive Bayes

The Naive Bayes algorithm identifies the appropriate concept of an instance by calculating the probability of each of the possible concepts given the context that is used.

The concept with the highest probability given the test vector is then assigned to the instance.

Naive Bayes is a probabilistic classifier that is based on the application of Bayes' Theorem and the assumption that the features in the feature vector are conditionally independent from each other. Bayes' Theorem is defined in Equation 2.5 for any event A and B.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{2.5}$$

The goal of the classifier in WSD is to determine the probability of each of the possible concepts, C, given an instance represented as a test vector, $\mathbf{F}$, which is written mathematically as $P(C|\mathbf{F})$. It is impossible though to observe every possible combination of features and concepts in the training data in order to calculate this.

Using Bayes Theorem, $P(C|\mathbf{F})$ is reformulated to Equation 2.6.

$$P(C|\mathbf{F}) = \frac{P(C)P(\mathbf{F}|C)}{P(\mathbf{F})} \tag{2.6}$$

The question of *given a particular vector what is the most likely concept* becomes *given a particular concept what is the most the likely vector.* This is still difficult to calculate. $P(\mathbf{F}|C)$ is calculated by dividing the number of times vector $\mathbf{F}$ is seen in the training data assigned to concept C by the number of times $\mathbf{F}$ occurs in the training data. The chances of $\mathbf{F}$ being seen in the training data at all is small due to the sparseness of the data.

This can be alleviated using the conditional independence assumption which assumes that the features in a feature vector are conditionally independent. This allows for the probability of a feature to be calculated independent of any of the other features. Rather than having to calculate the number of times the entire feature vector occurs with a concept in the training data, it is estimated based on observing the number of times an individual feature occurs with a concept as seen in Equation 2.7.

$$\frac{P(C)P(\mathbf{F}|C)}{P(\mathbf{F})} = \frac{P(C)\prod_{i}^{n} P(F_i|C)}{\prod_{i}^{n} P(F_i)} \tag{2.7}$$

The probability of a feature given a concept ($P(F_i|C)$) is calculated by dividing the number of times the feature is seen with the concept by the number of times the feature

occurs in the training data, the probability of a concept, $P(C)$, is calculated by dividing the number of times an instance in the training data is assigned that concept by the number of instances, and the probability of a feature $P(F_i)$ is calculated by dividing the number of times a feature occurs in the training data by the total number of features. The probability feature vector $P(F_i)$ is constant for all of the possible concepts and does not influence the outcome since the goal is to determine the concept with the maximum probability. The probabilities are often very small so log probabilities are used to prevent underflow. CD Therefore, the Naive Bayes model can be written so that concept C' is assigned to an instance such that:

$$\begin{aligned} C' &= \arg\max_C \frac{P(C)\prod_i^n P(F_i|C)}{\prod_i^n P(F_i)} \\ &= \arg\max_C P(C)\prod_i^n P(F_i|C) \\ &= \arg\max_C \log(P(C)\prod_i^n P(F_i|C)) \end{aligned}$$

For example, consider the following instance:

The control group consisted of a cold flush with heparinzed solution

where the word *cold* could refer to either the *Common Cold* (CC) or the *Cold Temperature* (CT) with the probability cold being assigned the concept *Common Cold* is 90% and *Cold Temperature* is 10% in our dataset.

The score for *Common Cold* (C^{CC}) and *Cold Temperature* (C^{CT}) using the Naive Bayes model and the probabilities of the features seen in Table 2.1 are calculated as follows:

Table 2.1: Naive Bayes Probability Table

Surrounding Word	P(Word|CT)	P(Word|CC)
control (c)	.95	.05
group (g)	.99	.01
consisted (cd)	.4	.6
flush (f)	.9	.1
heparinzed (h)	.5	.5
solution (s)	.5	.5

$$
\begin{aligned}
C^{CT} &= \log(P(CT)\prod_{i}^{n} P(F_i|CT)) \\
&= \log(P(CT)(P(\mathrm{c}|CT)P(\mathrm{g}|CT)P(\mathrm{cd}|CT)P(\mathrm{f}|CT)P(\mathrm{h}|CT)P(\mathrm{s}|CT))) \\
&= \log(.90 * (.95 * .99 * .40 * .90 * .50 * 0.50)) \\
&= \log(.1523) \\
&= -1.1183
\end{aligned}
$$

$$
\begin{aligned}
C^{CC} &= \log(P(CC)\prod_{i}^{n} P(F_i|CC)) \\
&= \log(P(CC)(P(\mathrm{c}|CC)P(\mathrm{g}|CC)P(\mathrm{cd}|CC)P(\mathrm{f}|CC)P(\mathrm{h}|CC))P(\mathrm{s}|CC))) \\
&= \log(0.1 * (0.05 * 0.01 * 0.60 * 0.10 * 0.50 * 0.50)) \\
&= \log(0.0000015) \\
&= -6.1249
\end{aligned}
$$

The score for *Common Cold* ($C^{CC} = -6.1249$) which is greater than the score for *Cold Temperature* ($C^{CT} = -1.1183$) therefore the concept *Cold Temperature* is assigned to the target word.

2.2.2 Clustering WSD Methods

Clustering methods rely on unannotated training data. In general these methods perform word sense discrimination rather than disambiguation. Discrimination seeks to cluster instances of a given target word such that instances that use the same concept of the target word are in the same cluster, while disambiguation seeks to determine the appropriate concept of an instance given a sense-inventory. In order to evaluate clustering methods, though, the disambiguation of the words in a test data set is required.

One advantage to clustering is that a large amount of manually annotated training data is not required; the labeling of the instances in the training data is done using clustering algorithms, rather than by human annotators as with the supervised methods. Another advantage of clustering is that it is language and domain independent requiring only a corpus in the language and domain of interest. The disadvantage is that training data is required for each word that needs to be disambiguated and historically this method does not obtained as high of a disambiguation accuracy as supervised methods.

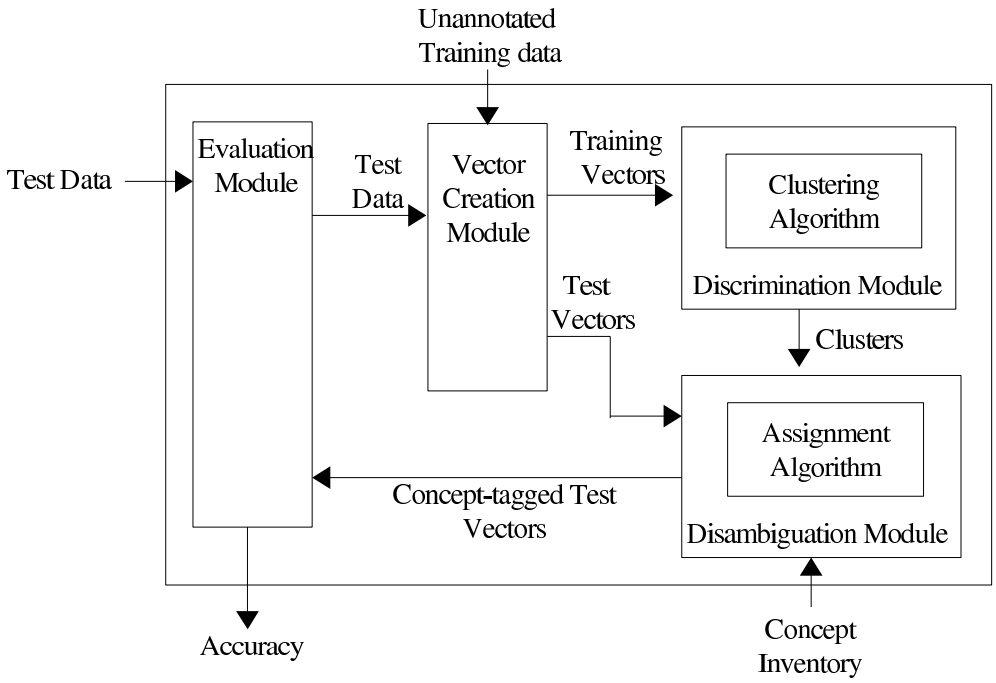

Figure 2.11: Clustering WSD Method

Figure 2.11 shows a general model of a clustering WSD method. First, test data is sent to the evaluation module. This data maybe annotated for evaluation purposes, if so, the concepts are removed and the data is sent to the vector creation module.

The vector creation program takes the test data and a set of unannotated training data that contains instances of the target word as input. A vector is created for each instance in the training and test data. The training vectors are sent to the discrimination module and the test vectors are sent to the disambiguation module.

In the discrimination module, the training vectors are grouped together using a clustering algorithm. A clustering algorithm plots the vectors in an n-dimensional space and groups them together into "clusters" as seen in Figure 2.12. There are a number of different types of clustering algorithms including agglomerative, divisive, and partitional algorithms. Agglomerative algorithms start with each training vector in a separate cluster and merge clusters, divisive methods start with all instance vectors in one cluster and split clusters, and partitional algorithms divide the training vectors into a preset number of clusters.

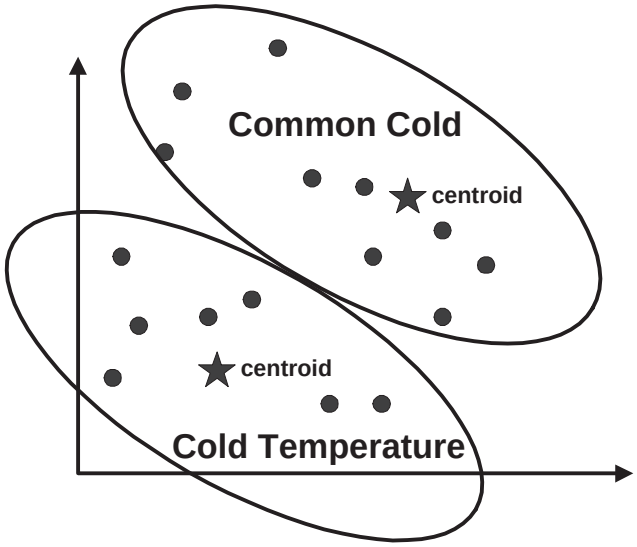

Figure 2.12: Clustering Example

The clusters are then sent to the disambiguation module where an assignment algorithm assigns a concept to each of the clusters using a sense-inventory. This can be done in a variety of different ways, one approach, described by [Wagstaff and Cardie, 2000], uses a small set of manually annotated training data to determine assignment of clusters. A concept vector is created for each possible concept by calculating the centroid of its cluster as denoted by the star in Figure 2.12. A test vector is then disambiguated by calculating the angle between it and each of the possible concept vectors using the cosine measure as seen in Figure 2.13. The concept whose vector is closest is assigned to the target word. This is done for each of the test vectors and then sent to the evaluation module to determine the accuracy of the system.

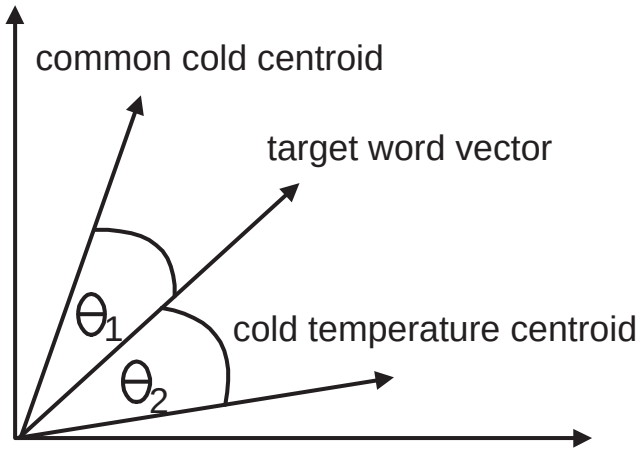

Figure 2.13: Assignment Algorithm Example

2.2.3 Knowledge-based Methods

Knowledge-based methods rely on information that can be extracted or inferred from a knowledge source, such as a dictionary, thesaurus or lexical database. These methods learn based on information from curated and structured data whereas supervised and clustering methods learn from example instances.

The advantage of the knowledge-based methods over the supervised and the clustering methods is that training data is not required for each word that needs to be disambiguated. This allows the system to disambiguate words in running text, referred to as

all-words disambiguation. All-words disambiguation methods have an advantage over what is termed *lexical-sample disambiguation* methods because lexical-sample methods can only disambiguate words in which there exists an ample set of training data. All-word disambiguation methods are scalable and can be used in real-word practical applications in which ambiguous words may not be known ahead of time and training data is difficult to obtain. The disadvantage to this method is that it is language and domain dependent because a knowledge source is required in the appropriate language and domain. Historically, it has also not obtained as high of a disambiguation accuracy as supervised methods.

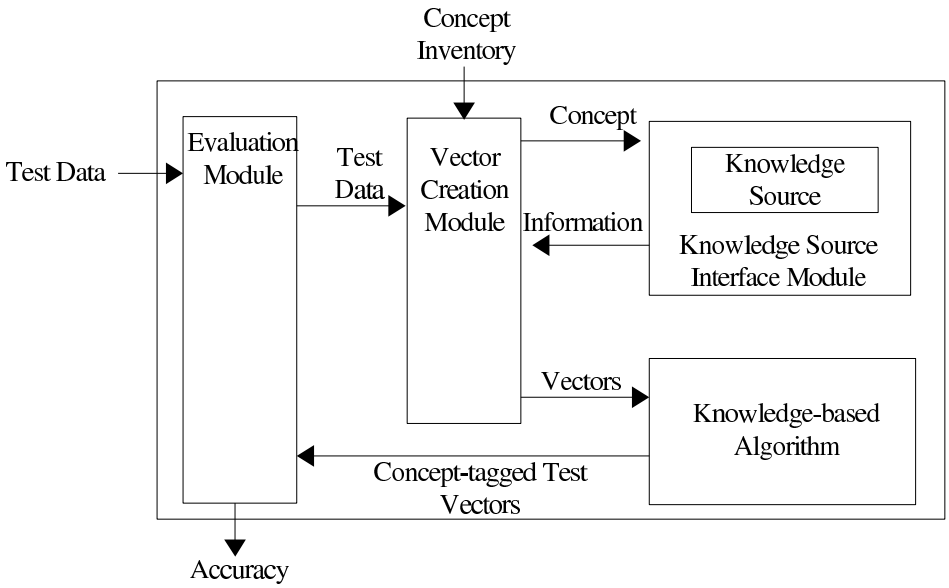

Figure 2.14: Knowledge-based WSD Method

Figure 2.14 shows a general model of knowledge-based WSD methods. In this method, the evaluation program takes the test data as input. The instances in the test data may be assigned their appropriate concept for evaluation purposes. These concepts are removed and the data is then sent to the vector creation module. A test vector is created for each instance in the test data using information from the knowledge source. This information is obtained through the knowledge source interface module. The information obtained and knowledge source used varies, for example, [Mohammad and Hirst, 2006] use the category information from Macquarie's machine

readable thesaurus, [Pedersen et al., 2005] use the semantic relatedness and similarity between the possible concepts and the words in the same context as the target word, and

The test vectors are then sent to the *knowledge-based algorithm*, which uses the information in the vectors to determine the appropriate concept of the target word. There are several different types of knowledge-based methods, but they all rely on human curated structured knowledge sources such as dictionaries, thesauri and/or lexical databases. The concept-tagged test data is then sent to the evaluation program and the accuracy of the method is returned.

The remainder of this section describes two different knowledge-based algorithms that have been used in WSD: a similarity algorithm and a vector algorithm.

Knowledge-based Similarity Algorithm

Semantic similarity and relatedness measures have been applied to the task of WSD in the general English domain. Semantic similarity and relatedness measures assign a score as to how similar or related two concepts are to each other. Semantic relatedness is a more general form of semantic similarity. For example, $foot$ and *sock* are related but not similar, where as $foot$ and *hand* are both similar and related. For more information about similarity and relatedness measures see Appendix A.

This method has previously been used to disambiguate words in general English using the knowledge source WordNet described in more detail in Section 2.3.1. [Banerjee and Pedersen, 2003], [Altintas et al., 2005], and [Pedersen et al., 2005] use this algorithm to evaluate various semantic similarity and relatedness measures.

In this method, for each instance in the test data a concept vector is created for each possible concept. So if there are two possible concepts, two concept vectors are created for each instance in the test data. The features set consists of the words in the test instance and the elements are the similarity score between the feature and the concept. These vectors are passed to the knowledge-based algorithm which calculates the average of the similarity scores for each of the vectors. The concept with the highest similarity score is assigned to the instance.

For example, there are two possible concepts of the target word cold: $Cold\ Temperature$ and $Common\ Cold$. Consider the following test instance:

The control group consisted of a cold flush with heparinzed solution.

A concept vector is created for *Cold Temperature* and *Common Cold* where the features are the following content words from the test instance:

- control
- group
- consisted
- flush
- heparinzed
- solution

The elements in the *Cold Temperature* concept vector are the semantic similarity scores between the concept *Cold Temperature* and each of the features as seen in Figure 2.15. The elements in the *Common Cold* concept vector are the semantic similarity or relatedness s cores between the concept *Common Cold* and its feature as seen in Figure 2.16.

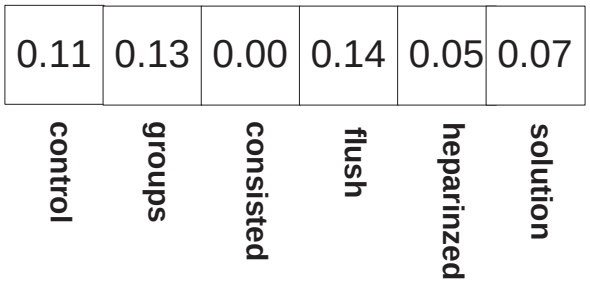

Figure 2.15: Similarity Vector for Cold Temperature

The elements in the *Cold Temperature* concept vector are summed and divided by the number of features obtaining an overall score of 0.083, and, similarly, the elements in the *Common Cold* concept vector are summed and divided by the number of features obtaining an overall score of 0.042. The target word is assigned the concept *Cold Temperature* because it has the highest overall score.

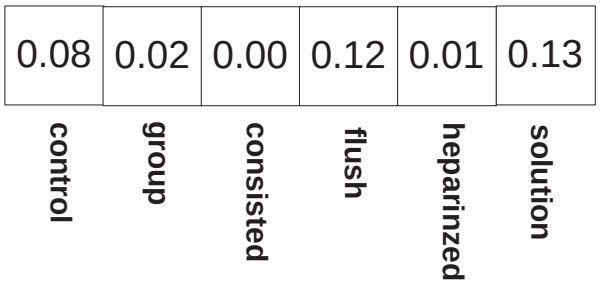

Figure 2.16: Similarity Vector for Common Cold

Knowledge-based Vector Algorithm

In this method, the vector creation module creates a test vector for each instance in the test data and a concept vector for each possible concept of the target word. The concept vector is created using information about that concept from a knowledge source such as its definition or synonyms terms, for example, [Patwardhan, 2003] use the definitions of a concept and its related concepts, and [Mohammad and Hirst, 2006] and [Humphrey et al., 2006] use the terms in a knowledge source associated with a concepts categorization.

The knowledge-based algorithm takes the vectors as inputs and a measure such as the cosine, dice or Euclidean distance is used to quantify the distance between the test vector and each of the possible concept vectors in an n-dimensional space. The concept whose vector is closest to the test vector is assigned to the target word.

There exists a range of classification methods that use the location of a vector in some n-dimensional space to determine its class such as this and the clustering method described in Section 2.2.2.

The clustering of feature vectors goes back at least as far as the beginning of Information Retrieval (IR) research, for example, [Salton et al., 1975] proposed a clustering method to automatically index documents for retrieval. This approach treats documents as vectors and clusters them in an n-dimensional space. A new document vector is compared to each of the clusters centroid to determine into which cluster it should be placed. This basic method is cited as the foundation for the clustering word sense discrimination method proposed by [Schütze, 1992].

The classification of feature vectors in some n-dimensional space also dates back at

least as far as the 1950 where [Fix and Hodges, 1951] proposed a supervised learning algorithm called k-Nearest Neighbor, determines the classification of a test vector based the class of its k closest training vectors where "closeness" is calculated using the Euclidean distance metric. Recently, [Agirre and Martinez, 2004] introduce a supervised WSD methods that combine aspects of each of these methods. In this method, a concept vector is created for each possible concepts of the target word by calculating the centroid of manually annotated training vectors assigned that concept. A test vector is assigned a concept by calculating the angle between it and each of the concept vectors using the cosine measure. The concept whose vector is closest to the test vector is assigned to the target word.

The main difference between these different methods is the training data used and how the concept vector is created. In the supervised and clustering methods a concept vector is created using the centroid of training vectors assigned that concept; manually annotated in the case of supervised method and automatically annotated in the case of the clustering method. In the knowledge-based vector method, the concept vector is created using a context about the concept extracted from a knowledge-source.

The next section discuss the various knowledge-sources and training data that has been used with the WSD methods discussed in this chapter.

2.3 Knowledge Sources

A knowledge source is human curated data whose information is organized in a fixed structure. These sources are typically domain and language dependent and often contain hierarchical relations among the information.

There are many different types of knowledge sources; three commonly used ones are machine readable dictionaries such as Longman's Dictionary of Contemporary English (LDOCE), machine readable thesauri such as the Roget's Thesaurus, and lexical databases such as WordNet and the Unified Medical Language System (UMLS).

WordNet is a lexical database that contains general English concepts whereas the UMLS contains concepts from the biomedical and clinical domains. Each of the databases contain relation information between the concepts and now will be discussed in more detail.

2.3.1 WordNet

WordNet [Fellbaum, 1998] is a lexical database of English. WordNet 3.0 contains 155,287 words that are grouped together based on their synonymy. These groupings are called *synsets*. As of WordNet 3.0, there exists 117,659 synsets in WordNet. A word-synset pair is defined by its part-of-speech, definition, and a set of example sentences. WordNet contains terms that have one of the four possible parts-of-speech: noun, verbs, adjectives and adverbs.

The majority terms are nouns which comprise over 117,000 of the terms and 82,000 of the synsets. The majority of polysemous words are verbs followed by adjectives and then nouns. The average number of synsets for a verb is 2.17 while the average number of synsets for a Noun is 1.24. Table 2.2 shows the breakdown of the words, synsets and polysemy according to their part-of-speech.

Table 2.2: WordNet Statistics

Part-of-speech	# Word	# Synsets	# Word-Synset Pair	Average # Synsets
Noun	117,798	82,115	146,312	1.24
Verb	11,529	13,767	25,047	2.17
Adjective	21,479	18,156	30,002	1.40
Adverb	4,481	3,621	5,580	1.25
Total	155,287	117,659	206,941	1.52

Synsets are linked together through semantic relations such as: *hypernym*, *hyponym*, *meronym*, and *holonym*. 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_2) is a kind of vehicle (w_1), therefore a truck is a hyponym 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 relations between synsets occurs only within their respective part-of-speech. As mentioned above, a word-synset pair is defined by its part-of-speech, definition, and a

set of example sentences. Two word-synset pairs with different parts-of-speech would not have a semantic relation between them. This creates four distinct hierarchies within WordNet, one for each of the four parts-of-speech. These hierarchies are not connected; to link them together, a WordNet node must be created which links to the top most synset in each of the four hierarchies.

2.3.2 Unified Medical Language System

The Unified Medical Language System (UMLS) is a knowledge representation framework designed to support biomedical and clinical research. It includes over 100 knowledge sources and classification systems[1] encoded with different semantic and syntactic structures. The three major components of UMLS are the Metathesaurus, Semantic Network and SPECIALIST Lexicon.

The Metathesaurus is a multi-lingual lexical database that combines information about biomedical and health-related concepts from various biomedical and clinical sources. The sources that have been semi-automatically integrated into the UMLS[2] . Some sources include: National Cancer Institute Thesaurus (NCI), SNOMED Clinical Terms (SNOMED-CT), and Medical Subject Headings (MSH).

The NCI Thesaurus is a biomedical terminology database developed and maintained by the National Cancer Institute[3] . It contains 1,300,000 concepts mapped to 4,600,000 terms with 17,000,000 relations. SNOMED-CT is an extensive clinical terminology developed by the College of American Pathologists (CAP) and maintained by the International Health Terminology Standards Development Organisation (IHTSDO). SNOMED-CT contains 315,000 concepts with formal logic-based definitions and is organized in a hierarchical structure. SNOMED-CT is the largest source in the UMLS Metathesaurus. MSH is the National Library of Medicine's controlled vocabulary thesaurus. MSH contains 25,186 concepts, each containing a set of associated terms used to describe it. The concepts are arranged in a hierarchical structure.

The Metathesaurus organizes knowledge based on Concept Unique Identifiers (CUIs). In the UMLS version 2009AB, there exists approximately 1.5 million CUIs.

[1] For a complete listing of sources:http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html
[2] A full source listing can be found at: http://www.nlm.nih.gov/research/umls/metaa1.html
[3] http://ncimeta.nci.nih.gov/MetaServlet/

A CUI is expressed by having specific attributes that define it such as its:

- preferred term
- concept definition
- associated terms
- related concepts

For example, the CUI C0009264 has the preferred term *Cold Temperature* in the 2008AB version of the UMLS. In the remainder of this dissertation, the preferred term of a CUI is often used with the actually CUI in brackets next to it for clarity. The definition of Cold Temperature Cold Temperature [C0009264] is:

> Having less heat energy than the object against
> which it is compared; the absence of heat

Some of the terms associated with Cold Temperature [C0009264] are:

- Cold Temperature
- Low Temperature
- Cold Thermal Agent

These are terms that are commonly used to describe the CUI and include the preferred term in its list. The are 12 different types of relations that can exist between concepts:

- PAR/CHD: parent/child
- RB/RN: broader/narrower than
- SY: source asserted synonymy
- RO: has a relationship other than synonymous, narrower, or broader
- RL: concepts are similar or "alike".
- RQ: related and possibly synonymous
- SIB: sibling
- AQ: allowed qualifier
- QB: can be qualified by
- RQ: related and possibly synonymous
- RU: related but unspecified
- XR: not related

Not all CUIs have all of the above relations and two concepts may have more than one relation between them. The relations used in this dissertation are PAR/CHD, RB/RN, SY, and SIB. Examples of CUIs that have the above relations with Cold Temperature [C0009264] can be seen in Table 2.3.

Table 2.3: Relations of Cold Temperature [C0009264] in the UMLS

PAR	Thermal Agent [C0542315] Temperature [C0039476] Weather [C0043085]
CHD	Freezing [C0016701] Extreme Cold [C1830748]
RB	Temperature [C0039476]
RN	None
SY	None
SIB	Rain [C0034640] Snows [C0037386] Light Emitted by the Sun [C0038817] Temperature[C0039476] Wind, NOS [C0043187] Transition Temperatures [C1257885] Temperatures, Hot [C2350229]

The information about a CUI comes from the source information. A concept within a specific source is called an Atom Unique Identifier (AUI). For example, the AUI Cold [A12785313] is from NCI and the AUI Low Temperature [A3292554] is from SNOMED-CT. The AUIs from the sources are semi-automatically combined to form CUIs. Cold [A12785313] from NCI and Low Temperature [A3292554] from SNOMED-CT are both mapped to the Cold Temperature [C0009264].

All the attributes associated with an AUI are also associated with its corresponding CUI. For example, Common Cold [A0041261] from MSH and Common Cold [A0476539] from CRISP are mapped the Common Cold [C0009443] and each of them has their own corresponding definition which become the definition(s) of Common Cold [C0009443]:

- Common Cold [A0041261]: A catarrhal disorder of the upper respiratory tract, which may be viral or a mixed infection. It generally involves a runny nose, nasal

congestion, and sneezing. (from MSH)

- Common Cold [A0476539]: catarrhal disorder of the upper respiratory tract, which may be viral or a mixed infection; marked by acute coryza, slight rise in temperature, chilly sensations, and general indisposition. (from CRISP)

The relation information between CUIs also comes from the relation information at the AUI level and curated information from the UMLS editors. For example, in NCI there exists a relation between Cold [A12785313] and Temperature [A7574004]. The merging of the AUIs from different sources creates relations between the CUIs. Since, Cold [A12785313] maps to Cold Temperature [C0009264] and Temperature [A7574004] maps to Temperature [C0039476], the relation between [A12785313] and [A7574004] is mapped to the CUI level creating a relation between [C0009264] and [C0039476]. For more in depth discussion about the relation information in the UMLS see Appendix B.

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 CUIs that are meaningfully related in some way. For example, the semantic type of Cold Temperature [C0009264] is assigned the semantic type "Natural Phenomenon or Process", where as Temperature [C0039476] is assigned the semantic type "Quantitative Concept". A CUI may be assigned more than one semantic type.

As of UMLS version 2009AB, there exist 135 semantic types. Other examples of semantic types include: Organism, Anatomical Structures, Biologic Function, and Chemicals. For a complete list of the semantic types see Appendix C.

The semantic types are connected by 54 semantic relations. Examples of semantic relations include: *is-a*, $part\text{-}of$, $ingredient\text{-}of$, $measurement\text{-}of$. For example, the semantic types "Quantitative Concept" and "Amino Acid Sequence" have the semantic relation $measurement\text{-}of$. For a complete list of the semantic relations see Appendix D.

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). There exists a lexical entry for each spelling or spelling variation.

2.4 Software Resources

This section discusses three software resources used in this dissertation:

- MetaMap
- WEKA data mining package
- SenseClusters

MetaMap is a concept mapping system which maps terms to concepts in the UMLS. Concept mapping is a general term in the biomedical domain which refers to the mapping of words and terms to concepts in an lexical database. Concept mapping systems were developed to aid in the retrieval and indexing of biomedical articles. MetaMap is used in the dissertation to extract biomedical information from text to be used as features.

The WEKA data-mining package developed by [Witten and Frank, 1999] contains software for data pre-processing, classification, regression, clustering, association rules, and visualization. This package implements some of the supervised learning algorithms such as those described in the supervised WSD methods.

SenseClusters is a word sense discrimination package developed by [Purandare and Pedersen, 2004] and [Kulkarni and Pedersen, 2005]. The remainder of this section discusses these three software packages.

2.4.1 MetaMap

MetaMap is a concept mapping system that maps terms in biomedical text to concepts in the UMLS by identifying the UMLS CUIs of the content words in the text. The default version of MetaMap does not perform word sense disambiguation. If an ambiguous term has more than one possible mapping, it returns all of them rather than disambiguating between them. However, in 2009 though, MetaMap was released with a WSD component. Prior to this, no WSD component was incorporated into MetaMap. This section first describes the original implementation of MetaMap and then discusses the WSD component.

Figure 2.17 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.

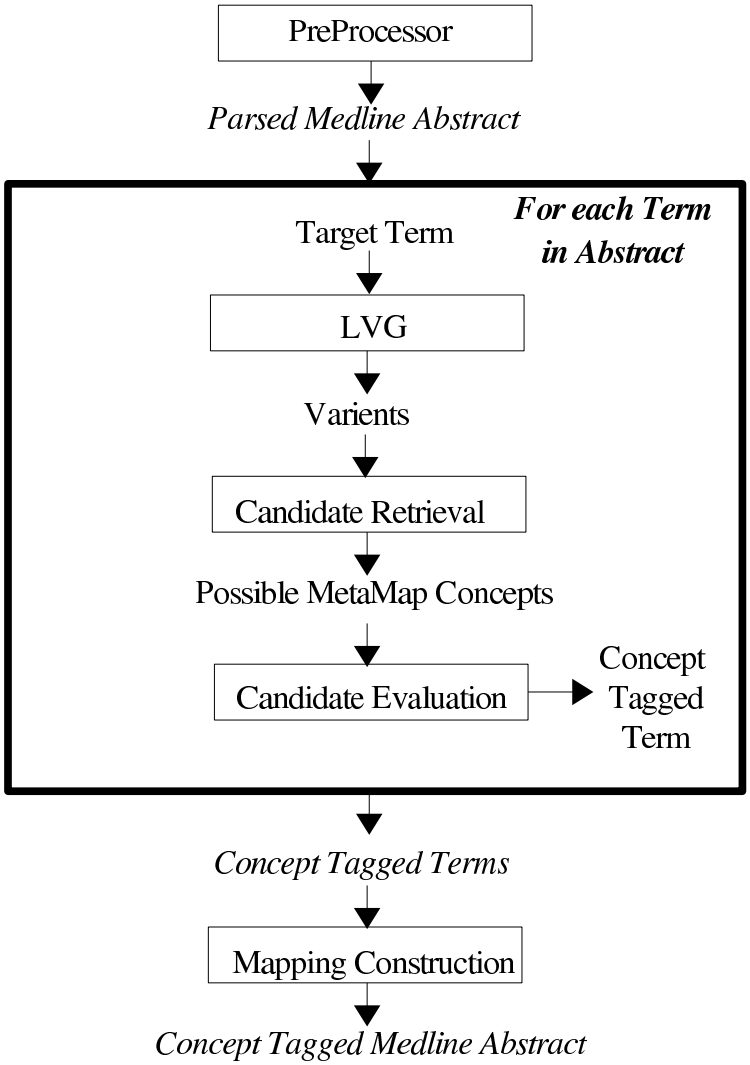

Figure 2.17: Current MetaMap System

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. The concept(s) with the highest MTI score are assigned to the associated term. This score based on four criteria:

- centrality
- variation
- coverage
- cohesiveness

Centrality is whether the potential concepts preferred term contains the head of the input data term. *Variation* is the distance between the input data term and potential concept preferred term. [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 preferred term. 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 concepts preferred term. For example, for the term "inferior vena cava stent filter" and the potential concept "Vena Cava Filter" have two words the consecutively overlap.

The MTI score, used to determine which of the candidate concepts should be chosen, is not based on the context in which the term is used but the form of the term itself. The criteria of centrality, variation, coverage and cohesiveness are quantifying the similarity between the term itself and the preferred term of the possible concept but does not take the context in which the word is used into consideration. This differs from WSD which does determine the appropriate concept based on the context that it is being used.

This does not imply that MetaMap can never map an ambiguous word to the correct CUI. MetaMap does not map individual words to concepts in the UMLS but terms. Consider the single word term *culture*. MetaMap maps this term to two possible CUIs in the UMLS: Anthropological Culture [C0010453] and Laboratory Culture [C0430400]. Now consider the multi-word term *laboratory culture*. MetaMap maps this term to only the CUI Laboratory Culture [C0430400] because of the coverage and variation criteria components used to calculate the MTI score for each of the possible concepts.

In 2009, MetaMap was released with a WSD component. This component is an implementation of the WSD knowledge-based system proposed by [Humphrey et al., 2006] which is discussed in Chapter 7. Using this component, when a term is returned by MetaMap with multiple possible CUIs, it is sent to the WSD component which disambiguates the term and returns the appropriate CUI. This CUI is then mapped to the term rather than all of them.

The information provided by MetaMap has been used as features to create feature vectors for WSD methods such as the supervised WSD method proposed by [Leroy and Rindflesch, 2004] whose feature set contains the semantic types of the words in the same sentence as the target word. Given an instance, MetaMap provides the following information:

- phrasal information
- part-of-speech (POS) information
- terms in the sentence
- CUIs of the terms
- semantic types of the CUIs

Consider the following sentence:

The groups susceptibility to a cold appeared to be positively associated with the risk.

MetaMap splits the sentence into eight phrasal units with the POS of each of the terms:

- the (determiner) groups (noun) susceptibility (noun)
- to (preposition) a (determiner) cold (noun) appeared to (adverb)
- be (auxiliary)
- positively (adverb)

- and (conjunction)
- associated (verb)
- cool (verb)
- with (preposition) the (determiner) risk (noun)

MetaMap then assigns a CUI with its associated semantic type seen in Table 2.4. Notice that MetaMap mapped the term *common cold* to the correct CUI because of the coverage and variation criteria components used to calculate the MTI score for each of the possible concepts.

Table 2.4: MetaMapped Terms

Term	CUI	Semantic Type
groups	Groups [C0441833]	Idea or Concept
susceptibility	Susceptibility [C1547045]	Quantitative Concept
	Susceptibility [C0012655]	Disease susceptibility
	Predisposition [C0220898]	Organism Attribute
	Susceptibility [C1264642]	Functional Concept
cold	Common Cold [C0009443]	Disease or Syndrome
	Cold Temperature [C0009264]	Natural Phenomenon or Process
	Cold Sensation [C0234192]	Physiologic Function
associated	Associated with [C0332281]	Qualitative Concept
risk	Risk [C0035647]	Qualitative Concept

MetaMap (without the WSD component) is used in this dissertation to obtain the CUIs of the terms in the same instance as the target word in the proposed supervised WSD method discussed in Chapter 3.

2.4.2 WEKA Data Mining Package

The WEKA data-mining package is a Java package containing tools for data preprocessing, classification, regression, clustering, association rules, and visualization. WEKA stands for "Waikato Environment for Knowledge Analysis" and was developed at the University of Waikato in New Zealand. WEKA contains a number of supervised learning algorithms such as Support Vector Machines (SVMs) and the Naive Bayes algorithm.

This package can be used to implement a supervised WSD method which were described above. In the supervised WSD method, a training vector is created for each instance in a set of manually annotated training data. Each instance in the training data is manually assigned the appropriate concept of the target word. The training vectors are taken as input by a supervised learning algorithm which creates a model based on the information in the vectors. A test vector is then created for each instance in the test data. The model then assigns the test vectors their appropriate concept.

The input format required by WEKA for the test and training vectors is called ARFF format. The format for these files must be exactly the same. The only difference between them is the vectors themselves.

For example, consider the following training instances which are sentences from the NLM-WSD dataset:

- He used a combination of the UW solution both for initial flush and the *cold* immersion.
- The groups susceptibility to a *cold* appeared to be positively associated with the risk.

and the following test instance:

- The control group consisted of *cold* flush with heparinzed saline.

where *cold* is the target word and has three possible concepts: $Cold\ Temperature$, $Common\ Cold$ and $None$.

Figure 2.18 shows a small example of the training vectors in ARFF format where the instance numbers are denoted after the % in the last two lines. In this example, the features are the words surrounding the target word in the training data and the part-of-speech of the target word.

In the ARFF format, the "@RELATION" tag identifies what the ARFF file contains. In our example:

@RELATION cold

indicates that the dataset is for the target word *cold*. The dataset contains 26 features and one classification label denoted by the "@ATTRIBUTE" tag. The last ATTRIBUTE is always the classification of the instance. In this example, there are referring to the three possible concepts. Instance one (%1) is assigned the concept $Cold$

```
@RELATION cold
@ATTRIBUTE he NUMERIC
@ATTRIBUTE used NUMERIC
@ATTRIBUTE a NUMERIC
@ATTRIBUTE combination NUMERIC
@ATTRIBUTE of NUMERIC
@ATTRIBUTE the NUMERIC
@ATTRIBUTE UW NUMERIC
@ATTRIBUTE solution NUMERIC
@ATTRIBUTE both NUMERIC
@ATTRIBUTE for NUMERIC
@ATTRIBUTE initial NUMERIC
@ATTRIBUTE flush NUMERIC
@ATTRIBUTE and NUMERIC
@ATTRIBUTE cold NUMERIC
@ATTRIBUTE immersion NUMERIC
@ATTRIBUTE groups NUMERIC
@ATTRIBUTE susceptibility NUMERIC
@ATTRIBUTE to NUMERIC
@ATTRIBUTE appeared NUMERIC
@ATTRIBUTE be NUMERIC
@ATTRIBUTE positively NUMERIC
@ATTRIBUTE associated NUMERIC
@ATTRIBUTE with NUMERIC
@ATTRIBUTE risk NUMERIC
@ATTRIBUTE POS {adjective, adverb, noun, verb}
@ATTRIBUTE Concept {Cold_Temperature,Common Cold,None}
@DATA
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,adjective,Cold_Temperature %1
0,0,1,0,0,1,0,0,0,0,0,0,0,1,0,1,1,1,1,1,1,1,1,1,noun,Common Cold           %2
```

Figure 2.18: Training Vectors in ARFF Format

Temperature and instance two (%2) is assigned the concept *Common Cold*.

The first 24 features are defined by a numeric attribute. In this case, a one or a zero indicating the existence of the feature in the instance. For example, the first instance, indicated by %1 after the "@DATA" tag, contains the first feature "he" indicated by a 1 in the first element position. The second instance (%2) does not contain the feature "he" indicated by a zero in this same position. The last feature, "POS", is defined by four nominal attributes: adjective, adverb, noun and verb. In the first instance the target word *cold* is an adjective whereas in the second it is a noun.

Figure 2.19 shows an example of a test vector in ARFF format. The format for the test and training vectors is exactly the same. The feature "he" is the first feature in the training file as well as the test file. In the test data the concept of the target word is not given but is denoted by a "?". The model assigns the test vector the appropriate concept.

```
@RELATION cold
@ATTRIBUTE he NUMERIC
@ATTRIBUTE used NUMERIC
@ATTRIBUTE a NUMERIC
@ATTRIBUTE combination NUMERIC
@ATTRIBUTE of NUMERIC
@ATTRIBUTE the NUMERIC
@ATTRIBUTE UW NUMERIC
@ATTRIBUTE solution NUMERIC
@ATTRIBUTE both NUMERIC
@ATTRIBUTE for NUMERIC
@ATTRIBUTE initial NUMERIC
@ATTRIBUTE flush NUMERIC
@ATTRIBUTE and NUMERIC
@ATTRIBUTE cold NUMERIC
@ATTRIBUTE immersion NUMERIC
@ATTRIBUTE groups NUMERIC
@ATTRIBUTE susceptibility NUMERIC
@ATTRIBUTE to NUMERIC
@ATTRIBUTE appeared NUMERIC
@ATTRIBUTE be NUMERIC
@ATTRIBUTE positively NUMERIC
@ATTRIBUTE associated NUMERIC
@ATTRIBUTE with NUMERIC
@ATTRIBUTE risk NUMERIC
@ATTRIBUTE POS {adjective, adverb, noun, verb}
@ATTRIBUTE Concept {Common_Cold,Cold_Temperature,None}
@DATA
0,0,0,0,1,1,0,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,1,0,adjective,? %1
```

Figure 2.19: Test Vector in ARFF Format

This dissertation uses the WEKA data mining package in the proposed supervised WSD method discussed in the next chapter. WEKA offers flexibility by providing a variety of well tested supervised learning algorithms.

2.4.3 SenseClusters Package

SenseClusters[4] is a freely available open source PERL package that performs word sense discrimination.

In the SenseClusters, a training vector is created for each instance in an unannotated set of training data. These instances are grouped together by a clustering algorithm. SenseClusters uses the CLUTO software package[5] which is a computationally efficient clustering and cluster analysis package. SenseClusters uses the following clustering

[4] http://sourceforge.net/projects/senseclusters/
[5] http://glaros.dtc.umn.edu/gkhome/views/cluto

algorithms from CLUTO: Agglomerative, Graph partitional-based, Partitional biased agglomerative and Direct k-way clustering. The clustering can be done in either vector space where the vectors are clustered directly, or similarity space, where vectors are clustered by finding the pair-wise similarities among the contexts.

In the evaluation step, an assignment algorithm then assigns a concept to each of the clusters using a sense-inventory. The assignment algorithm uses a small set of manually annotated training data to determine assignment of clusters such as the approach. A concept vector is then created for each concept by calculating the centroid of its cluster. A new instance is then disambiguated by first creating a target word vector to represent this new instance. Then the angle is calculated between the test vector and each of the concept vectors using the cosine measure as seen above in Figure 2.13. The concept whose vector is closest is assigned to the target word.

SenseClusters represents the training, test and concepts vectors as either:

- first-order co-occurrence vectors
- first-order unigram vectors
- first-order bigram vectors
- second-order co-occurrence vectors
- second-order bigram vectors

Unigrams and bigrams are classified as *ngrams* which are defined as an ordered set of n words. For example, consider the example instance:

The control group consisted of cold flush with heparinzed solution.

The *unigrams* (1-grams) consist of the following content words:

- control
- group
- consisted
- cold
- flush
- heparinzed
- solution

The words *the*, *of* and *with* are not included. These words are considered stopwords. Stopwords are function words that do not contain information about the content of the instance. Lists of stopwords are typically manually compiled based on the domain of the task. An example of a stoplist can be seen in Appendix F. The *bigrams* (2-grams) consists of the following pairs of word:

- control group
- group consisted
- consisted cold
- cold flush
- flush heparinzed
- heparinzed solution

The co-occurrences are bigrams in which the order does not matter. For example, the bigram "control group" and "group control" are considered two different features. With co-occurrence they would be considered a single feature.

This dissertation uses the SenseClusters programs that create the first and second-order vectors in the proposed knowledge-based WSD method that will be discussed in Chapter 5

2.5 Unannotated Data

This section discusses the unannotated datasets that have been used by WSD Methods in the biomedical domain and general English and referred to in this dissertation. The biomedical dataset is called Medline[6] . The general English datasets are: the Brown corpus[7] the British National Corpus[8] , and the Wall Street Journal Corpus[9] The remainder of this section describes the four datasets.

2.5.1 Medline

Medline is an abbreviation for "Medical Literature Analysis and Retrieval System Online". It is a bibliographic database containing over 16 million citations to journal

[6] http://mbr.nlm.nih.gov/
[7] http://khnt.aksis.uib.no/icame/manuals/brown/
[8] http://www.natcorp.ox.ac.uk/
[9] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42

articles in the biomedical domain which is maintained by NLM. The citations come from approximately 5,200 journals in 37 different languages starting from 1949. The majority of the publications are scholarly journals but a small number of newspapers, magazines, and newsletters have been included. MEDLINE is the primary component of PubMed[10] which is a free online repository which allows access to Medline as well as other citations and abstracts in the fields of medicine, nursing, dentistry, veterinary medicine, health care systems, and pre-clinical sciences. An example of a Medline citation in PubMed is as follows:

PMID- 9378719
TI - A reassessment of the molecular origin of cold denaturation.
AB - The existence of cold denaturation is now firmly demonstrated by its direct observation for several globular proteins inaqueous solution. But the physico-chemical explanation of this intriguing phenomenon is still unsatisfactory. In this paper we deepen our understanding of cold denaturation by taking advantage of the theoretical model developed by Ikegami and using thermodynamic data on the transfer to water of liquid N alkyl amides. The analysis leads to the conclusion that the presence of water is fundamental to determine the existence of cold denaturation due to its strong energetic interaction with the amino acid residues previously buried in the protein's interior.

The PMID (PubMed Identifier) refers to the reference number of the citation. TI refers to the title of the citation and AB refers to its abstract. Not all citations have an associated abstract but they all have a title and reference number.

This dissertation uses the 2005 Medline baseline[11] in its experiments. The baseline contains 14,792,864 citations in Medline dating from 1949 to 2005. Each citation was processed using MetaMap. The baseline contains 2,043,918 unique words (tokens) and 295,585 unique CUIS.

[10] http://www.ncbi.nlm.nih.gov/sites/entrez
[11] http://skr.nlm.nih.gov/resource/MetaMappedBaselineInfo.shtml

2.5.2 The Brown Corpus

The Brown corpus was created by H. Kucera and W. N. Francis at Brown University, Providence, RI. The corpus contains 1,014,312 words and consists of 500 samples of text published in the United States in the year 1961. The samples are distributed across 15 genres including news articles, editorials, reviews, religious text, periodicals, government documents, non-fiction and fiction books. Each sample contains 2,000 or more words.

2.5.3 The British National Corpus

The British National Corpus (BNC) was created and is maintained by the BNC Consortium led by Oxford University Press. The corpus contains 100 million words and consists of samples of written and spoken English from a wide range of sources including newspapers, specialist periodicals, journals, academic books and popular fiction.

2.5.4 The Wall Street Journal Corpus

The Wall Street Journal Corpus (WSJ) is a subset of the Penn Treebank[12]. The corpus contains 2,499 stories from a three year Wall Street Journal (WSJ) collection of 98,732 stories for syntactic annotation. Each story has been annotated with its syntactic and semantic information.

2.6 Concept-Tagged Data

This section discusses the datasets created for WSD in the biomedical and general English domains. These datasets contain instances of specific target words that have been manually annotated. These data sets are used to evaluate WSD methods. Due to the supervised methods requirement of manual annotation, the datasets are often split into a training and test portions.

[12] http://www.cis.upenn.edu/ treebank/

2.6.1 Biomedical Dataset

NLM-WSD Dataset

National Library of Medicine's Word Sense Disambiguation (NLM-WSD) dataset contains top 50 most frequent ambiguous words from the 1998 Medline baseline. Each target word in the NLM-WSD dataset 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 concept in the UMLS (CUI) or assigned the concept as "None" if none of the possible concepts described the term.

Appendix E shows the possible CUIs with their preferred term in the UMLS for each of the target words in the dataset. Each instance has also been processed by MetaMap providing phrasal, part-of-speech, CUI and semantic type information for each of the terms surrounding the target word.

The NLM-WSD dataset is currently the only freely available biomedical dataset created specifically for word sense disambiguation where each instance containing an ambiguous word is assigned a concept from the UMLS.

The disadvantage of the dataset is that it is small, only 100 instances for each of the target words and there is a very high *majority sense baseline* for many of the target words. The majority sense baseline is the accuracy that would be achieved if all the instances were assigned to the concept with the greatest number of instances. There are 15 out of the 50 terms whose majority sense is less than 65%. Table 2.5 shows the majority sense baseline for each of the target words in the data set. The table also contains the number of instances assigned a concept or None for each of the target words. A blank space indicates that the target word does not have that concept. For example, there are only three possible concepts of the target word *adjustment*.

There exists some target words in which very few concepts exist in the sense-inventory. For example, all of the instances for the target word *association* are tagged with *None* meaning there does not exist a UMLS concept to describe the concept for any of the instances. Another example is the target word *fluid* which only contains instances that have been assigned Concept 1.

Unfortunately, the dataset itself does not contain the actual CUI of the possible concept but rather the CUIs preferred term. The preferred term of a CUI can change over time therefore the actual CUI is required for the methods proposed in this dissertation. In order to retrieve the actual CUIs, an exact look up was conducted using the MRCON table in the 1999 version of the UMLS. The MRCON table in the UMLS contains a list of all of the possible CUIs in the UMLS along with their preferred term. A complete list of this CUIs can be seen in Appendix E.

2.6.2 General English Datasets

"interest", "line", "hard", and "serve" Datasets

The "interest" dataset [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 concepts from the Longman Dictionary of Contemporary English (LDOCE).

The "line" dataset [Leacock et al., 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 concepts from WordNet.

The "hard" dataset [Leacock et al., 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 concepts from WordNet.

The "serve" dataset [Leacock et al., 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 concepts from WordNet.

SemCor Dataset

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 concept-tagged instances in the training data.

Table 2.5: NLM-WSD dataset

target word	Maj. concept (%)	Concept 1	Concept 2	Concept 3	Concept 4	Concept 5	None
adjustment	62	18	62	13			7
association	100	0					100
blood pressure	53	53	2	45			0
cold	86	86	6	1	0	2	5
condition	90	90	2				8
culture	89	11	89				0
degree	63	63	2				35
depression	85	85	0				15
determination	79	0	79				21
discharge	74	1	74				25
energy	99	1	99				0
evaluation	50	50	50				0
extraction	83	83	5				12
failure	71	4	25				71
fat	71	2	71				27
fit	82	0	18				82
fluid	100	100	0				0
frequency	94	94	0				6
ganglion	93	7	93				0
glucose	91	91	9				0
growth	63	37	63				0
immunosuppression	58	58	42				0
implantation	81	17	81				2
inhibition	98	1	98				1
japanese	74	5	74				21
lead	71	27	2				71
man	58	58	1	33			8
mole	83	83	1	0			16
mosaic	52	45	52	0			3
nutrition	45	45	16	28			11
pathology	85	14	85				1
pressure	96	96	0	0			4
radiation	60	60	38				2
reduction	89	2	9				89
repair	52	52	16				32
resistance	97	3	0				97
scale	65	0	65	0			35
secretion	99	1	99				0
sensitivity	49	49	1	1			49
sex	80	15	5	80			0
single	99	1	99				0
strains	92	1	92				7
support	90	8	2				90
surgery	98	2	98				0
transient	99	99	1				0
transport	93	93	1				6
ultrasound	84	84	16				0
variation	80	20	80				0
weight	47	24	29				47
white	49	41	49				10

Senseval Datasets

Senseval is an international organization There are whose goal is to promote research in WSD. The organization runs evaluation exercises to test WSD systems. There are four evaluations: SENSEVAL-1, which took place in 1998, SENSEVAL-2, which took place in 2001, SENSEVAL-3, which took place in 2004, and SENSEVAL-4, now called SEMEVAL-1, which took place in 2007. Currently, SEMEVAL-2 is scheduled to take place in 2010.

The SENSEVAL-1 dataset includes on English lexical sample which 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 dataset includes 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.

The SENSEVAL-3 dataset includes 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 datasets was WordNet1.7 for nouns and WordSmith for verbs.

The SENSEVAL-4 (also called the SEMEVAL-1) datasets includes English-Chinese parallel text, Turkish lexical sample, English all-words and English lexical sample.

Chapter 3

K-CUI

This chapter discusses the proposed supervised WSD method called K-CUI which is designed to disambiguate words in biomedical text. This method uses manually annotated training data mapped by MetaMap to *learn* the context in which the target words are used. K-CUI's feature set contains the CUIs of the terms surrounding the target word in the training data. The supervised learning algorithm creates a model using these features and automatically assigns concepts to instances in the test data. The novelty of K-CUI is using MetaMap to map terms to CUIs in the UMLS to be used as features in a supervised WSD method.

The following sections discuss the motivation of using CUIs as features, the algorithm used to implement K-CUI and then the actual K-CUI implementation.

3.1 Motivation

K-CUI uses the CUIs mapped by MetaMap to terms in the training data as features in its feature set. The motivation behind using CUIs is the assumption that CUIs provide term level information that may not always be captured within the term itself. For example, each of the following terms map to the CUI Falls [C0085639]:

- fall
- falls
- falls down
- falling

- falling down
- fell down
- fell

When using the individual terms as features, each term is considered a different feature, but, when using the CUI of a term, all of the individual terms correspond to a single feature. Consider the following two instances:

- He *fell down* the stairs and broke his arm.
- She kept *falling* regardless of the medication.

When using terms as features, *fell down* and *falling* are considered two different features hence both of these instances contain distinct features. When using the CUIs of the terms rather than the terms itself both instances contain the feature Falls [C0085639].

Not all of the CUIs included in the feature set are useful for disambiguation. For example, the feature set for the target word *cold* contain the following CUIs that seemingly have nothing to do with any of the possible concepts of *cold*:

- Effective
- Control
- Plants
- Compare
- Three
- Role
- Mutant

This introduces noise into the feature set which is defined to be features that do not help distinguish between the different concepts of a target word. K-CUI has a two techniques to determine which CUIs surrounding the target word should be included in the feature set in order to reduce the amount of noise. The first technique is called *windowing* which determines the size of the window around the target word in which the CUIs can be extracted. A *window* is the number of words on either side of the target word.

The second technique uses a cutoff score to determine which CUIs to include in the feature set. K-CUI has three cutoffs options: a frequency cutoff, a MetaMap Indexing (MMI) cutoff and a semantic similarity cutoff. When using one of the cutoff options, a

feature is included in the feature set only if it obtains a score above a specified threshold. This reduces the number of features in the feature set by removing those features that are not relevant to the disambiguation process.

The remainder of this chapter discusses the algorithm describing the proposed supervised method, the actual implementation of the method, and then the windowing and cutoff options.

3.2 Algorithm

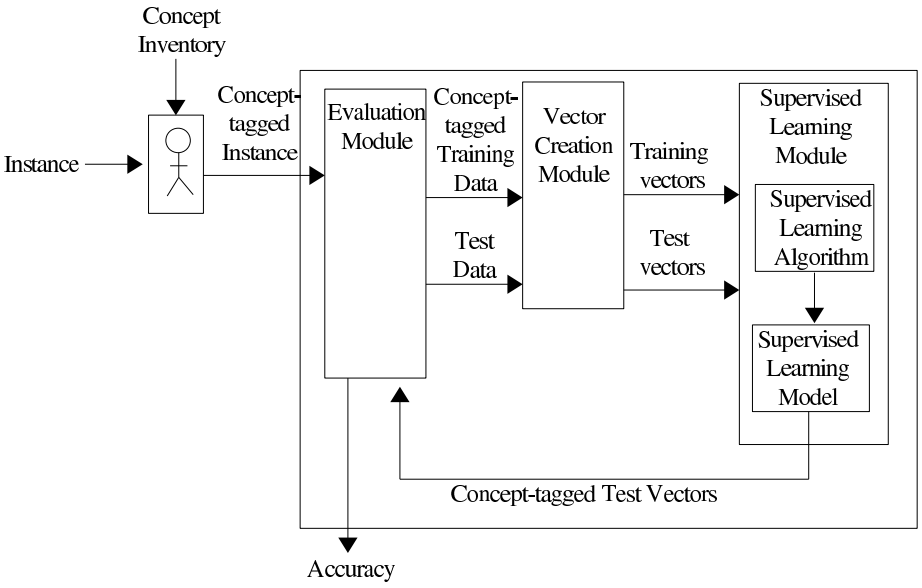

Figure 3.1: K-CUI Algorithm

This section describes the K-CUI algorithm shown in Figure 3.1. K-CUI is composed of three modules: the evaluation module, the vector creation module, and the supervised learning module. The evaluation module sets up the testing framework in order to determine the accuracy of the K-CUI experiments which is necessary for evaluation purposes. The vector creation module extracts the features from the training data and creates the test and training vectors. The supervised learning module creates the learning model which assigns concepts to the instances in the test data. The pseudocode for the main driver program of K-CUI is in Algorithm 3.1.

Algorithm 3.1 K-CUI Pseudocode

```
procedure K-CUI(ManuallyAnnotatedData)
  comment: Randomly Split Data Into Data into X blocks
  Blocks = SplitData(ManuallyAnnotatedData, X)
  for each Block ∈ Blocks
  {
    comment: Step 1: Create Test and Training Data
    TestData = RemoveAnnotations(Block)
    TrainingData = MergeBlocks(Blocks, Block)
    comment: Step 2: Create Training and Test Vectors
    (TrainingVectors, TestVectors) = CreateVectors(TrainingData, TestData)
    comment: Step 3: Create Learning Model
    SupervisedLearningModel = CreateSupervisedLearningModel(TrainingVectors)
    comment: Step 4: Assign Concepts to Test Vectors
    ConceptTaggedTestVectors = AssignConcepts(TestVectors, SupervisedLearningModel)
    comment: Step 5: Calculate Accuracy of the System
    Accuracy = CalculateAccuracy(ConceptTaggedTestVectors, AnnotatedData)
    TotalAccuracy = TotalAccuracy + Accuracy
  }
  TotalAccuracy = TotalAccuracy/NumberOfBlocks
  printTotalAccuracy
```

As the pseudocode of the main K-CUI drivers shows, there are five main steps. In step 1, the *Evaluation Module* creates the training and test data; the pseudocode for this module is shown in Algorithm 3.2. The SplitData() function takes the manually annotated data as input and splits the data into X blocks to perform X-fold cross validation. In X-fold cross validation, the instances are divided into X blocks where each block contains an equal number of instances. The model is created using the X-1 blocks as training data and then tested using the remaining block. This is repeated X times such that each block has been used as test data exactly once with the remaining blocks used as training data. The accuracy is calculated at each fold and the average accuracy is returned.

Algorithm 3.2 Evaluation Module Pseudocode

function SplitData($ManuallyAnnotatedData$, X)
comment: Split Data into X blocks
$Blocks$ = Split($ManuallyAnnotatedData$, X)
return ($Blocks$)

function RemoveAnnotations($Data$)
comment: Remove Annotations (Concepts) from the Dataset
$UnannotatedData$ = RemoveConcepts($Data$)
return ($UnannotatedData$)

function MergeBlocks($Blocks$, $Block$)
for each $B \in Blocks$
if $B! = Block$
then ConcatenateBlock(B, $MergedBlock$)
return ($MergedBlock$)

function CalculateAccuracy($ConceptTaggedTestVectors$, $AnnotatedData$)

$Correct$ = GetNumberCorrect($ConceptTaggedTestVectors$, $AnnotatedData$)
$Wrong$ = GetNumberWrong($ConceptTaggedTestVectors$, $AnnotatedData$)
$Accuracy = Correct$ / ($Correct + Wrong$)
return ($Accuracy$)

Algorithm 3.3 Vector Creation Module Pseudocode

```
function CreateVectors(TrainingData, TestData)
 FeatureSet = CreateFeatureSet(TrainingData)
 TrainingVectors = CreateFirstOrderVectors(FeatureSet, TrainingData)
 TestVectors = CreateFirstOrderVectors(FeatureSet, TestData)
 return (TrainingVectors, TestVectors)

function CreateFeatureSet(TrainingData)
 comment: Process Data using MetaMap and extract the CUIs
 MetaMappedData = MetaMap(TrainingData)
 FeatureSet = ExtractCUIs(MetaMappedData)
 return (FeatureSet)

function CreateFirstOrderVectors(FeatureSet, Data)
 MetaMappedData = MetaMap(Data)
 for each Instance ∈ MetaMappedData
 {  comment: Create vector where each element is a feature in the Feature Set
    Vector = InitializeVector(Vector)
    for each Feature ∈ Vector
    {  if Feature ∈ Instance
       then Vector[Feature] = 1
       else Vector[Feature] = 0
    comment: Add the Vector to an array of Vectors to be returned
    Vectors ← Vector
 return (Vectors)
```

In step 2, the *Vector Creation Module* creates the training and test vectors; the pseudocode for this module is shown in Algorithm 3.3. In this module, the CreateVectors() function takes the training and test data as input and creates the feature set by calling the CreateFeatureSet() function which extracts the CUIs assigned by MetaMap to instances in the training data. It then calls the CreateFirstOrderVectors() function which creates the first-order training and test vectors using the features from the feature set.

Algorithm 3.4 Supervised Learning Module Pseudocode

```
function CreateSupervisedLearningModel(TrainingVectors)
  SupervisedLearningModel = SupervisedLearningAlgorithm(TrainingVectors)
  return (SupervisedLearningModel)

function AssignConcepts(TestVectors, SupervisedLearningModel)
  ConceptTaggedTestVectors = SupervisedLearningModel(TestVectors)
  return (ConceptTaggedTestVectors)
```

In step 3, the *Supervised Learning Module* creates a supervised learning model, and, then in step 4, assigns concepts to the instances in the test data using this model; the pseudocode for this module is shown in Algorithm 3.4. The CreateSupervisedLearningModel() takes the training vectors as input and creates a supervised learning module using a supervised learning algorithm such as the Naive Bayes. The AssignConcepts() function takes the model and the test vectors as input and assigns a concept to each of the vectors.

In step 5, the *Evaluation Module* calculates the accuracy of the model; the pseudocode for this module is shown in Algorithm 3.2. The CalculateAccuracy function takes the concept tagged test vectors and the manually assigned data as input and calculates the accuracy of the assignments. The following section describes the actual implementation details of this algorithm.

3.3 System

This section discusses the implementation details of K-CUI which is shown in Figure 3.2. K-CUI takes instances of a specific target word that have been manually assigned a concept from the sense-inventory as input, for example, Figure 3.3 shows an example of ten instances containing the target word *cold* where each instance has been manually assigned one of two concepts: Cold Temperature and Common Cold. The instances are in 'plain' text and each instance is annotated with its target word, concept and instance id. The evaluation module randomly splits these instance into a training and test dataset for evaluation purposes. The module then removes the concepts assigned to instances in the test and sends both of the datasets to the vector creation module.

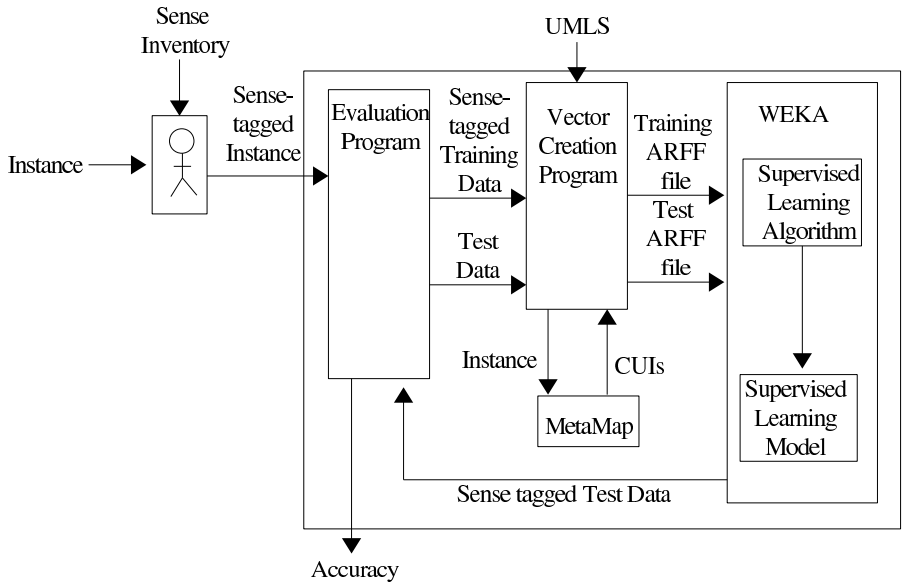

Figure 3.2: K-CUI System

The vector creation module creates a first-order feature vector for each of the training and test instances using the CUIs of the terms surrounding the target word in the training data as features. These CUIs are obtained using MetaMap which takes the training and test instance as input and returns them with each term mapped to one or more CUIs in the UMLS if there exists a mapping. A feature set is created using the CUIs mapped to the terms surrounding the target word in the training data. The

```
The results suggest that in smooth muscle induction of NO pathway
relaxation, which is in part mediated by K+ channels and inducible NO
synthase, may be of importance to the understanding of
ischemia/reperfusion responses in <head item="cold" instance="1"
sense="Cold_Temperature>cold</head> stored arteries.

We constructed a model of peripheral nerve messages in an attempt to
represent and quantitate the desynchronizations produced by <head
item="cold" instance="2" sense="Cold_Temperature>cold</head> and crush
damage lesions in peripheral nerve messages.

In this paper we deepen our understanding of <head item="cold"
instance="3" sense="Cold_Temperature>cold</head> denaturation by taking
advantage of the theoretical model developed by Ikegami and using
thermodynamic data on the transfer to water of liquid N-alkyl amides.

Thus, the scintigraphy pattern of a hot spot in the bone scan and a
<head item="cold" instance="4" sense="None">cold</head> lesion in the
bone marrow scintigraphy is highly suggestive of a mandibular
metastasis, if accompanied by anesthesia of the lower lip.

Two-dimensional gel analysis showed that protein synthesis was
deregulated in csp double mutants and that the loss of one or two CSPs
led to an increase in the synthesis of the remaining CSP(s) at 37
degrees C and after <head item="cold" instance="5"
sense="Cold_Temperature>cold</head> shock, suggesting that CSPs down-
regulate production of members from this protein family.

The control group consists of (<head item="cold" instance="6"
sense="Cold_Temperature>cold</head> flush with heparinized saline.

Group I comprised 9 dogs submitted to renal autotransplantation and
group II comprised 6 dogs submitted to renal autotransplantation after
24 h <head item="cold" instance="7" sense="Cold_Temperature>cold</head>
ischemia.

Personal histories of hypertension and thyroid disease, and
susceptibility to <head item="cold" instance="8"
sense="Common_Cold">cold</head>s appeared to be positively associated
with the risk.

TIR1/SRP1 has previously been identified as a gene induced by glucose,
<head item="cold" instance="9" sense="Cold_Temperature>cold</head>
shock or anaerobiosis and was believed to be a cell membrane protein
but not a cell wall protein.

Relationship between <head item="cold" instance="10"
sense="Cold_Temperature>cold</head> tolerance and generation of
suppressor macrophages during acute stress.
```

Figure 3.3: Instances in Plain Text

module then creates a training vector for each instance in the training data and a test vector for each instance in the test data. The vectors contain the CUIs from the feature set and the elements are either a one or a zero indicating whether or not the CUI occurs in the instance. Figures 3.4 and Figure 3.5 show an example of the training and the test vectors created by the vector module in the ARFF format required by WEKA.

```
@RELATION cold
@ATTRIBUTE C0009264 NUMERIC %Cold Temperature
@ATTRIBUTE C0009443 NUMERIC %Common Cold
@ATTRIBUTE C0234192 NUMERIC %Cold Sensation
@ATTRIBUTE C0205263 NUMERIC %Induced
@ATTRIBUTE C1563921 NUMERIC %Cold Ischemia
@ATTRIBUTE C1442770 NUMERIC %24h
@ATTRIBUTE C0031119 NUMERIC %Peripheral Nerves
@ATTRIBUTE C0038659 NUMERIC %Suggestion
@ATTRIBUTE C0470166 NUMERIC %Message
@ATTRIBUTE C1280551 NUMERIC %Dog Family
@ATTRIBUTE C1280200 NUMERIC %Entire Peripheral Nerve
@ATTRIBUTE C1518425 NUMERIC %Not Otherwise Specified
@ATTRIBUTE C0162340 NUMERIC %Comprehension
@ATTRIBUTE C1515023 NUMERIC %Submitted
@ATTRIBUTE C0036974 NUMERIC %Shock
@ATTRIBUTE C0012984 NUMERIC %Canis Familiaris
@ATTRIBUTE C0221198 NUMERIC %Lesion
@ATTRIBUTE C0194194 NUMERIC %Autotransplantation of kidney
@ATTRIBUTE Concept {Cold Temperature,Common Cold,None}
@DATA
0,0,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,Cold Temperature % 2
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,Common Cold      % 8
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,Cold Temperature % 9
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,Cold Temperature % 10
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,None            % 4
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,Cold Temperature % 3
0,0,0,0,1,1,0,0,0,1,0,0,0,1,0,1,0,1,Cold Temperature % 7
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,Cold Temperature % 5
0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,Cold Temperature % 1
```

Figure 3.4: Training Vectors in ARFF Format

In this example, nine out of the ten instances from Figure 3.3 are used as training data while the remaining instance is used as test data. In the ARFF files, the "@RELATION" tag indicates that the dataset is for the target word *cold* and the CUI after the "@ATTRIBUTE" tag is a feature from the feature set except for the last one labelled

```
@RELATION cold
@ATTRIBUTE C0009264 NUMERIC %Cold Temperature
@ATTRIBUTE C0009443 NUMERIC %Common Cold
@ATTRIBUTE C0234192 NUMERIC %Cold Sensation
@ATTRIBUTE C0205263 NUMERIC %Induced
@ATTRIBUTE C1563921 NUMERIC %Cold Ischemia
@ATTRIBUTE C1442770 NUMERIC %24h
@ATTRIBUTE C0031119 NUMERIC %Peripheral Nerves
@ATTRIBUTE C0038659 NUMERIC %Suggestion
@ATTRIBUTE C0470166 NUMERIC %Message
@ATTRIBUTE C1280551 NUMERIC %Dog Family
@ATTRIBUTE C1280200 NUMERIC %Entire Peripheral Nerve
@ATTRIBUTE C1518425 NUMERIC %Not Otherwise Specified
@ATTRIBUTE C0162340 NUMERIC %Comprehension
@ATTRIBUTE C1515023 NUMERIC %Submitted
@ATTRIBUTE C0036974 NUMERIC %Shock
@ATTRIBUTE C0012984 NUMERIC %Canis Familiaris
@ATTRIBUTE C0221198 NUMERIC %Lesion
@ATTRIBUTE C0194194 NUMERIC %Autotransplantation of kidney
@ATTRIBUTE Concept {Cold_Temperature,Common_Cold,None}
@DATA
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,? % 6
```

Figure 3.5: Test Vector in ARFF Format

"Concept" which indicates the possible concepts of the target word. In this case, there exist three possible concepts: *Cold Temperature*, *Common Cold* and *None*. The vectors, whose instance id's are denoted after the %, come after the "@DATA" tag. The vectors consist of zero and ones that indicate if the feature exists in its corresponding instance. For example, in Figure 3.4, the first training vector is instance two which is assigned the concept *Common Cold* and whose vector elements indicate that the following CUIs exist in its instance:

- Peripheral Nerves [C0031119]
- Messages [C0470166]
- Entire Peripheral Nerve [C1280200]

In Figure 3.5, the test vector is instance six whose vector elements indicate that the following CUIs exists in it its instance:

- Cold Temperature [C0009264]
- Common Cold [C0009443]

- Cold Sensation [C0234192]

The vectors are sent to the WEKA data mining package which uses a specified supervised learning algorithm to create a model that assigns each of the instance in the test vector a concept. The concept-tagged test vectors are returned to the evaluation module which calculates the accuracy of the module.

The remaining section discusses the different options that can be used with K-CUI to determine which CUIs to include in the feature set.

3.3.1 Windowing Options

Windowing refers to the location of the words surrounding the target word in which features are extracted. K-CUI has three windowing options available: phrase, sentence and abstract. The *phrase* option includes in the feature set only those CUIs that exist in the same phrase as the target word, the *sentence* option includes only those CUIs that exist in the same sentence as the target word, and the *abstract* option includes only those CUIs that exist in the same abstract.

For example, consider the following abstract by [Graziano et al., 1997] in which the terms have been mapped to CUIs by MetaMap:

The existence of *cold* denaturation [C0301642] is now demonstrated [C1999141] by its direct [C1947931] observation [C1964257] for several globular_proteins [C0178663] in aqueous [C0599956] solution [C0037633]. In this paper [C1547566], we deepen our understanding [C0162340] of denaturation [C0301642] by taking advantage of the theoretical_model [C0026350] developed [C1999145] by Ikegami and using [C1524063] thermodynamic [C0039808] data [C1511726] on the transfer to water [C1550678] of liquid [C0302908] N-alkyl [C1177206] amides [C000248261]. The analysis [C1524024] leads [C0181586] to the conclusion [C1707478] that the presence [C0392148] of water [C1550678] is fundamental to determine [C2004162] the existence of denaturation [C0301642].

The terms, identified by MetaMap, contain an underscore between the individual words in the term and the CUIs are in brackets to the right of term.

Using the phrase option, the CUIs are only extracted from the phrase "of *cold*

denaturation" therefore only the CUI [C0301642] is included in the feature set. Using the sentence option, the CUIs are only extracted from the sentence "The existence of *cold* denaturation is now demonstrated by its direct observation for several globular proteins in aqueous solution", therefore, only the CUIs [C0301642], [C1999141], [C1947931], [C1964257], [C0178663], [C0599956] and [C0037633] are included in the feature set. Using the abstract option, all of the above CUIs are included in the feature set.

3.3.2 Frequency Cutoffs

K-CUI contains the option of using a frequency cutoff to determine whether a CUI should be included in the feature set. The assumption is that CUIs that occur more often with the target word are a better indicator of the context in which a word is used than those that occur infrequently.

Consider the following instance of the term *cold*:

The control group consists of *cold* flush with heparinzed saline

Table 3.1 shows the CUIs surrounding the target word *cold* and the number of times the CUI is seen with *cold* in the manually annotated training data.

Table 3.1: Frequency Cutoff

CUI	Frequency
Control Group [C0009932]	9
Flushing [C0016382]	12
Saline [C0036082]	13

The CUIs included in the feature set are those whose frequency count is greater than the specified cutoff. For example, given a frequency cutoff of ten, a CUI that is seen ten or more times in the manually annotated training data would be included. In this example, only the CUI Control Group [C0009932] is excluded from the feature set because it only occurs nine times with the target word in the training data.

3.3.3 MMI Score Cutoff

K-CUI contains the option of using a MetaMap Indexing (MMI) cutoff to determine whether a CUI should be included in the feature set. An MMI score quantifies how

relevant a CUI is in describing a Medline abstract. A high MMI score indicates that the CUI is useful in describing the overall topic of the abstract. The assumption is that the more relevant a CUI is at describing the abstract the better it is able to distinguish between the possible concepts of a target word. For example, given the following abstract by [Graziano et al., 1997]:

The existence of cold denaturation is now firmly demonstrated by its direct observation for several globular proteins in aqueous solution. But the physico-chemical explanation of this intriguing phenomenon is still unsatisfactory. In this paper we deepen our understanding of cold denaturation by taking advantage of the theoretical model developed by Ikegami and using thermodynamic data on the transfer to water of liquid N-alkyl amides. The analysis leads to the conclusion that the presence of water is fundamental to determine the existence of cold denaturation due to its strong energetic interaction with the amino acid residues previously buried in the protein's interior.

Table 3.2 shows the ten CUIs assigned to the abstract by MetaMap. The first five are the CUIs assigned the highest MMI scores and the last five are the CUIs assigned the lowest MMI score. The scores indicate that the overall topic of the abstract is about Water [C0043047], Hydrotherapy [C0020311] and Observation in research rather than Strong [C0442821], ALKYL [C1177206] and Liquid [C1304698]. It is not that the low scoring CUIs do not exist in the abstract they are just not as central to the overall topic of the abstract.

An MMI score, proposed by [Aronson, 1997], was created to facilitate the indexing system called Medical Text Indexer (MTI). MTI recommends headings from the Medical Subject Headings (MSH) terminology to medical text indexers. The medical text indexers use these recommendations to manually assign a Medline citation one or more MSH headings for indexing purposes. The MSH headings exist in the UMLS as CUIs. The MMI score is used to help determine which MSH CUIs to recommend to the medical text indexers. MTI determines which MSH heading to recommend to the indexer based on the following steps:

- The citation is run through MetaMap, which returns a list of CUIs and their associated MetaMap Indexing (MMI) score.

Table 3.2: Top Five CUIs with the Lowest and Highest MMI Scores

CUI	MMI Score
Water [C0043047]	29
Hydrotherapy [C0020311]	23
Observation in research [C0302523]	16
Comprehension [C0162340]	15
Paper—C0030351]	15
Still [C1410088]	4
Chemicals [C0220806]	3
Strong [C0442821]	3
ALKYL [C1177206]	3
Liquid [C1304698]	3

- MTI then maps the CUIs that have a MMI score greater than 10 to MSH headings and recommends them to the indexers.

The MMI score for a given CUI is based on four components:

- the depth of the concept in the MSH hierarchy (d)
- the number of words in the concept (t)
- the number of characters in the concept (c)
- the frequency the concept occurs (f)
- the MetaMap score (m)

These components are based on the term associated with the concept and the location of the concept within the MSH hierarchy. The MMI score, like the MTI score discussed in Section 2.4.1, does not take into account the context in which the CUI or term is used in the calculation of the score. [Aronson, 1997] calculates the MMI score using the following formula:

$$mmi = v_f(f) \cdot \frac{w_d \cdot v_d(d) + w_t \cdot v_t(t) + w_c \cdot v_c(c) + w_m(m) \cdot v_m(m)}{w_m + w_w + w_c + w_m} \tag{3.1}$$

where

$$v_m(x) = ln(\frac{(e^m + 1) + (e^m - 1)x}{(e^m + 1) - (e^m - 1)x}/m) \tag{3.2}$$

and the coefficient w normalizes the components as follows:

- $w_f = f/10$
- $w_d = d/9$
- $w_t = t/26$
- $w_c = c/102$
- $w_{mm} = mm/1000$

Table 3.3: MMI Score Cutoff

CUI	MMI Score
Cold Temperature [C0009264]	2.2
Control Groups [C0009932]	1.8
Saline Solution [C0036082]	1.6
Common Cold [C0009443]	1.3
Flushing [C0016382]	1.0
Flush C1696091]	0.4
Cold Sensation [C0234192]	0.4

K-CUI uses the MMI score to determine which CUIs should be used as features. Table 3.3 shows the CUIs with their associated MMI score returned by MetaMap for the instance:

The control group consists of cold flush with heparinzed saline

Given a MMI score cutoff of 1.0, only the top five CUIs in the table would be used as features; the CUIs Flush [C169091] and Cold Sensation [C0234192] would not be included.

Semantic Similarity Cutoff

A semantic similarity cutoff is used in K-CUI to determine whether a CUI should be included in the feature set. Semantic similarity measures[1] quantify how similar two concepts are by determining their closeness in a hierarchy. The assumption behind using a similarity measure as a cutoff is that words that are used in the same context have a similar meaning, therefore, CUIs with a high similarity score are a better able

[1] Appendix A discusses semantic similarity measures in more detail

to distinguish between the possible concepts of a target word than CUIs with a low similarity score.

Consider two possible concepts for the target word *cold*: Cold Temperature [C0009264] and Common Cold [C0009443]. The concept Temperature [C0039476] has a semantic similarity score of 0.1250 with Cold Temperature [C0009264] and 0.0833 with Common Cold [C0009443] indicating that it would be an indicative feature, whereas the concept Benzoate [C0220795] has a semantic similarity score of 0.0833 with Cold Temperature [C0009264] and 0.0625 with Common Cold [C0009443] indicating that it would not be a good indicator as to which concept is being referred to.

In this option, the semantic similarity is calculated between the CUIs surrounding the target word and each of the possible concepts. If one of the scores is higher than the specified threshold, the CUI is included in the feature set. For example, consider the following instance containing the target word *cold*:

The control group consists of *cold* flush with heparinzed saline

where *cold* has the following possible concepts:

- Cold Temperature [C0009264]
- Common Cold [C0009443]

The semantic similarity score is obtained between each of the concepts of *cold* and the CUIs mapped to the terms in the instance. Table 3.4 shows the CUIs surrounding the target word *cold* and the maximum similarity score obtained between the CUI and each of the possible concepts.

Table 3.4: Similarity Score Cutoff

CUI	Similarity Score
Control Group [C0009932]	0.11
Flushing [C0016382]	0.10
Saline [C0036082]	0.13

The CUIs included in the feature set would only be the ones that have a similarity score greater than the specified cutoff. For example, given a similarity cutoff of 0.10, only CUI Flushing [C0016382] would be excluded in the feature set.

The semantic similarity scores are obtained by K-CUI using the package created by [McInnes et al., 2009], called UMLS::Similarity[2] , which is a platform independent, freely available, open source Perl module created to calculate the semantic similarity between CUIs in the UMLS. The module takes two CUIs as input and returns their semantic similarity given a specified semantic similarity measure. As of version 0.17, UMLS::Similarity contains a simple path-based measure and the semantic similarity measures proposed by:

- [Rada et al., 1989]
- [Wu and Palmer, 1994]
- [Leacock and Chodorow, 1998]
- [Nguyen and Al-Mubaid, 2006]

These measures are accessed by K-CUI through an API provided by the package.

[2] UMLS::Similarity can be downloaded at http://search.cpan.org/dist/UMLS-Similarity/

Chapter 4

K-CUI Results

This section conducts six experiments using K-CUI and discusses their results. Section 4.1 investigates using various window sizes to determine which features to include in the feature set. The purpose of these experiments is to determine whether using a larger window size obtains a higher disambiguation accuracy than a smaller window size.

Section 4.2 investigates using a cutoff to determine whether a CUI should be included in a feature set. K-CUI has three cutoff options: a frequency cutoff, a MMI cutoff and a similarity cutoff. The frequency cutoff includes in the feature set only those CUIs that occur often with the target word. The assumption is that CUIs that occur more often with the target word are a better indicator of the context in which a word is used than those that occur infrequently. The MMI cutoff includes in the feature set only those CUIs that have a high MMI score. The MMI score indicates how relevant a CUI is at describing the overall topic of the abstract. The assumption is that the more relevant a CUI the better it is at distinguishing between the possible concepts of a target word. The similarity cutoff includes in the feature set only those CUIs that obtain a high semantic similarity score with one of the possible concepts of the target word. The assumption is that CUIs that have a similar meaning to one of the possible concepts of the target word will be a good indicator of a target word's possible concept.

Section 4.3 compares the biomedical feature CUIs with the general English feature unigrams and Section 4.4 compares the K-CUI results to the results reported by researchers that have evaluated their supervised WSD methods using the NLM-WSD

dataset.

Section 4.5 conducts a case analysis of target words that obtained a low disambiguation accuracy in the above experiments.

The experiments have the following commonalities between them. K-CUI uses the NLM-WSD dataset as its training and test data and reports the accuracy of the results using 10-fold cross validation. K-CUI uses the SVM and Naive Bayes algorithm from the WEKA data mining package as its machine learning algorithms for these experiments. The experiments compare each of the results to the majority sense baseline which is the accuracy that would be achieved by assigning every instance of the target word with the most frequent sense as assigned by the human evaluators. This chapter also calculates the statistical significance between the results using the pairwise t-test, which compares the accuracy of a target word from the results of one experiment with the accuracy of the same target word from the results of another experiment. The pairwise t-test tests if the sum of the change between the accuracy of the two experiments differs statistically significantly from zero.

4.1 Windowing Results

This section discusses the results of the windowing experiment. The purpose of these experiments is to determine whether using a larger window size returns a higher disambiguation accuracy when using biomedical features to disambiguate words in biomedical text compared to a smaller window size.

Humans only require a small window size around a target word to determine its appropriate concept. An experiment conducted by [Choueka and Lusignan, 1985] found that only two or three words were needed indicating that only the words closest to the target word are required for disambiguation. [Gale et al., 1992] found this was not the case for computers. They show that larger window sizes returned better results when disambiguating words in general English using general English features, as did [Joshi et al., 2005] who used general English features to disambiguate words in biomedical text. This lead to the question of whether the biomedical features closest to the target word a better indicator of its concept than those further away.

The hypothesis of this experiment is that the feature set containing CUIs extracted

Table 4.1: Windowing Results

		Naive Bayes			SVM		
	Baseline	phrase	sentence	abstract	phrase	sentence	abstract
adjustment	0.62	0.72	**0.73**	0.65	0.71	0.71	0.65
association	**1.00**	**1.00**	**1.00**	0.98	**1.00**	**1.00**	**1.00**
blood pressure	0.54	0.55	0.59	0.54	0.55	**0.62**	0.51
cold	0.86	**0.89**	0.88	**0.89**	**0.89**	0.88	0.88
condition	0.90	0.90	**0.92**	0.90	0.89	0.91	0.90
culture	0.89	0.89	0.88	**0.91**	0.89	0.89	0.89
degree	0.63	0.63	**0.83**	0.81	0.63	0.82	0.82
depression	**0.85**	**0.85**	**0.85**	0.83	**0.85**	**0.85**	0.84
determination	0.79	0.79	0.74	0.79	0.79	0.78	**0.84**
discharge	0.74	0.73	0.88	**0.93**	0.74	0.82	0.92
energy	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.50	0.41	0.60	0.69	0.40	0.71	**0.74**
extraction	0.82	0.82	**0.84**	**0.84**	0.83	0.82	**0.84**
failure	0.71	0.70	0.67	0.68	0.71	**0.72**	0.71
fat	0.71	0.81	0.80	0.74	**0.84**	0.83	0.78
fit	0.82	0.82	0.85	0.85	0.82	**0.86**	**0.86**
fluid	**1.00**	**1.00**	**1.00**	0.98	**1.00**	**1.00**	**1.00**
frequency	0.94	0.92	0.94	0.94	0.94	0.94	**0.95**
ganglion	0.93	0.93	**0.95**	0.94	0.93	**0.95**	0.93
glucose	0.91	**0.93**	0.90	0.90	0.91	0.90	0.91
growth	0.63	0.65	0.70	**0.72**	0.65	0.67	0.70
immunosuppression	0.59	0.63	0.73	0.80	0.63	0.76	**0.87**
implantation	0.81	0.81	0.87	**0.94**	0.81	0.85	0.92
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
japanese	0.73	0.82	0.79	0.78	**0.83**	0.78	0.78
lead	0.71	0.77	0.85	0.90	0.76	0.82	**0.93**
man	0.58	0.83	**0.88**	0.83	0.82	0.86	0.84
mole	0.83	**0.88**	0.82	0.84	0.87	0.86	0.86
mosaic	0.52	0.69	0.68	**0.78**	0.71	**0.78**	0.76
nutrition	0.45	**0.59**	0.55	0.47	**0.59**	0.45	0.42
pathology	**0.85**	**0.85**	0.83	**0.85**	**0.85**	0.82	0.83
pressure	**0.96**	**0.96**	**0.96**	0.95	**0.96**	**0.96**	**0.96**
radiation	0.61	0.68	0.75	0.84	0.67	0.66	**0.86**
reduction	**0.89**	0.87	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**
repair	0.52	0.54	0.73	**0.91**	0.54	0.72	0.84
resistance	**0.97**	**0.97**	**0.97**	0.96	**0.97**	**0.97**	**0.97**
scale	0.65	0.80	0.80	0.77	**0.82**	0.81	0.75
secretion	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
sensitivity	0.49	0.54	0.76	**0.92**	0.49	0.72	0.87
sex	0.80	0.84	0.79	**0.88**	0.82	0.84	**0.88**
single	**0.99**	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**
strains	**0.92**	0.88	0.91	**0.92**	**0.92**	**0.92**	**0.92**
support	**0.90**	0.89	0.87	**0.90**	0.89	**0.90**	**0.90**
surgery	**0.98**	**0.98**	0.96	0.94	**0.98**	**0.98**	**0.98**
transient	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
transport	**0.93**	**0.93**	0.92	**0.93**	**0.93**	**0.93**	**0.93**
ultrasound	0.84	0.86	0.84	0.84	**0.89**	0.84	0.85
variation	0.80	0.79	0.80	**0.88**	0.79	0.82	0.86
weight	0.47	0.54	0.72	0.74	0.40	**0.75**	0.71
white	0.49	0.70	0.76	**0.78**	0.67	0.69	0.74
Overall Accuracy	0.78	0.81	0.84	**0.85**	0.81	0.84	**0.85**

from a larger window results in a higher disambiguation accuracy than a feature set containing CUIs extracted from a smaller window.

To test the hypothesis, this section analyzes the results of using three different windows size in which to obtain the CUIs. The first window size, *phrase*, includes only those CUIs in the same phrase as the target word, the second window size, *sentence*, includes only those CUIs in the same sentence as the target word, and the third window size, *abstract*, includes only those CUIs in the same abstract as the target word. Table 4.1 shows the majority sense baseline and the results of these experiments using the SVM and Naive Bayes algorithm. Table 4.2 shows the statistical significance between the different results.

The results show that there is no difference between the Naive Bayes and SVM results and each of the experiments obtains a higher overall accuracy than the baseline. Using the CUIs in the same phrase as the target word returns an accuracy of 81%, using the CUIs in the same sentence returns an accuracy of 84% and using CUIs in the same abstract returns an accuracy of 85%. The results show that the overall accuracy increases as the window size increases and the increase in accuracy is statistically significant.

Table 4.2: P-values using the Pairwise T-test for Windowing Results

		Naive Bayes			SVM		
	Window	phrase	sentence	abstract	phrase	sentence	abstract
	baseline	0.0008	0.00003	0.00002	0.00171	0.00002	0.000001
Naive Bayes	phrase		0.00429	0.00257	0.26504	0.00496	0.00168
	sentence				0.00805	0.36050	
	abstract				0.00392	0.02011	0.40454
SVM	phrase					0.00922	0.00251
	sentence						

Although the larger window size obtains the highest overall accuracy, it is only four percentage points higher than using the CUIs in the same phrase as the target word and one percentage point higher than using the CUIs in the same sentence. Analysis shows that there are on average two words per phrase and using the CUIs in the same phrase results in a much smaller feature set than using the CUIs in the same sentence or abstract. The average number of features in the phrase experiment is 61.16 whereas the

average number of feature in sentence experiment is 726.17 and abstract experiment is 1,444.37, which indicates that a majority of the instances can be disambiguated by just looking at the two or three CUIs closest to the target word but in order to disambiguate the remaining instances a larger window size is required.

The overall conclusion of this experiment is that extracting biomedical features from larger window sizes obtains a higher disambiguation accuracy which is consistent with the results reported by [Joshi et al., 2005]. Although, a majority of the instances were disambiguated by looking at the phrase containing the target word.

4.2 Cutoff Results

The windowing results in Section 4.1 indicate that using all of the CUIs in the same abstract as the target word obtain the highest overall disambiguation accuracy (85%). It also results in the largest feature set, containing on average 1,444.37 CUIs. Not all of the CUIs in the feature set are useful, only a small number of them trigger a specific concept most have very little to do with the actual disambiguation. In addition, the CUIs in the feature set are obtained automatically using MetaMap which does not attempt to disambiguate terms that map to more than one CUI, but instead returns all of the possible CUIs. K-CUI includes all of these CUIs in the feature set. This introduces noise into the feature set which is defined to be features that do not help distinguish between the different concepts of a target word. The question arose whether there is a way to alleviate some of the noise that may exist in the feature set without degrading the results.

To reduce the amount of noise in the feature set, K-CUI experiments with three cutoffs options. The first is the frequency cutoff, which only includes CUIs that occur with the target word at least a specified number of times. The second is the MMI cutoff, which only includes CUIs that have a high MMI score. The third is the similarity cutoff, which only includes CUIs that have a high similarity score with one of the possible concepts.

This section, first shows the results using each of the different cutoffs. Then discusses the overall results of using a cutoff in general.

4.2.1 Frequency Cutoff Results

This section discusses the results of the frequency cutoff experiment. The hypothesis of this experiment is that using a frequency cutoff will reduce the amount of noise in the feature set increasing the overall disambiguation accuracy. The assumption behind using a frequency cutoff is that CUIs that occur more often with the target word are a better indicator of the context in which a word is used than those that occur infrequently.

To test this hypothesis, this section analyzes the results of using a frequency cutoff of zero, two, four and six; a cutoff of zero is equivalent to not using a cutoff at all. Table 4.3 shows the majority sense baseline and the results for these experiments using the SVM and Naive Bayes algorithm. Table 4.4 shows the statistical significance between the results.

The results show that the Naive Bayes obtains an overall accuracy of 85%, 84%, 79% and 74% when using a cutoff of zero, two, four and six while the SVM obtains an overall accuracy of 85%, 84%, 81%, and 77%. There exists no difference in the overall accuracy between the algorithms when using a cutoff of zero or two. As the frequency cutoff increases, the Naive Bayes results degrade much quicker than the SVMs and the difference between the results is statistically significant.

The results using a frequency cutoff indicate that CUIs that occur only a few times in the training data are playing a significant role in the disambiguation process. Table 4.5 shows the average number of features in the feature set (Avg. # Features in Feature Set), the average number of features that occur in the training data and in the test instance (Avg. # Non-Zero Elements), and the overall disambiguation accuracy when using the Naive Bayes and SVM algorithms. The average number of features in the feature set when not using a cutoff is 2455.48, which decreases to 1426.98 when using a cutoff of two. Therefore, there exists approximately 1028.71 features that only occur once in the training data. The average number of features that occur in the training data and in the test instance is 50.77 when not using a cutoff, and 44.92 when using a cutoff of two therefore there exists approximately 5.85 features that only occur once in the training data and in the test data. The overall results decrease by only one percentage point using a frequency cutoff of two but analysis of the individual results shows the decrease can be much greater, for example, the accuracy for the target word *lead* decreases from 90% to 83% when the low frequency features are removed.

Table 4.3: Frequency Cutoff Results

		Naive Bayes				SVM			
	Baseline	No cutoff	2	4	6	No cutoff	2	4	6
adjustment	0.62	0.65	0.65	0.49	0.36	0.65	**0.67**	0.60	0.43
association	**1.00**	0.98	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
blood pressure	**0.54**	**0.54**	0.45	0.51	0.50	0.51	0.40	0.50	0.53
cold	0.86	0.89	**0.90**	0.86	0.79	0.88	0.88	0.88	0.87
condition	**0.90**	**0.90**	0.89	0.88	0.76	**0.90**	0.89	0.89	0.82
culture	0.89	0.91	**0.92**	0.88	0.89	0.89	0.90	0.88	0.90
degree	0.63	0.81	0.81	0.66	0.64	**0.82**	0.78	0.73	0.66
depression	0.85	0.83	0.81	**0.87**	0.83	0.84	0.84	0.83	0.85
determination	0.79	0.79	0.82	0.53	0.47	**0.84**	0.76	0.74	0.57
discharge	0.74	0.93	0.94	**0.95**	0.86	0.92	0.94	0.86	0.83
energy	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.50	0.69	0.65	0.64	0.66	**0.74**	0.63	0.68	0.62
extraction	0.82	0.84	**0.86**	0.84	0.81	0.84	0.83	0.80	0.53
failure	**0.71**	0.68	0.68	0.58	0.43	**0.71**	0.70	0.67	0.65
fat	0.71	0.74	0.73	0.67	0.59	**0.78**	0.76	0.73	0.69
fit	0.82	0.85	**0.86**	0.70	0.56	**0.86**	0.85	0.82	0.61
fluid	**1.00**	0.98	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
frequency	0.94	0.94	**0.95**	0.76	0.58	**0.95**	**0.95**	**0.95**	0.79
ganglion	0.93	0.94	**0.97**	0.92	0.90	0.93	0.95	0.96	0.95
glucose	**0.91**	0.90	0.90	0.87	0.90	**0.91**	0.88	0.84	0.89
growth	0.63	0.72	**0.73**	0.65	0.64	0.70	0.66	0.63	0.59
immunosuppression	0.59	0.80	0.79	0.77	0.79	**0.87**	0.76	0.76	0.81
implantation	0.81	**0.94**	0.91	0.79	0.62	0.92	0.92	0.91	0.81
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	0.95
japanese	0.73	**0.78**	0.75	0.64	0.53	**0.78**	0.73	0.75	0.70
lead	0.71	0.90	0.83	0.74	0.69	**0.93**	0.92	0.86	0.78
man	0.58	0.83	0.80	0.72	0.42	**0.84**	0.83	0.75	0.66
mole	0.83	0.84	0.85	0.69	0.82	**0.86**	0.85	0.81	0.75
mosaic	0.52	0.78	0.77	0.69	0.69	0.76	0.75	0.75	**0.80**
nutrition	0.45	**0.47**	0.42	0.34	0.32	0.42	0.38	0.42	0.33
pathology	**0.85**	**0.85**	0.73	0.59	0.64	0.83	0.84	0.72	0.61
pressure	**0.96**	0.95	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	0.89	**0.96**
radiation	0.61	0.84	0.82	0.75	0.71	**0.86**	0.84	0.71	0.69
reduction	0.89	0.89	0.89	0.86	0.89	0.89	0.89	0.83	**0.91**
repair	0.52	**0.91**	0.88	0.84	0.80	0.84	0.83	0.80	0.77
resistance	**0.97**	0.96	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	0.96	0.95
scale	0.65	0.77	0.77	**0.79**	0.67	0.75	0.74	0.73	**0.79**
secretion	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.98
sensitivity	0.49	**0.92**	0.88	0.83	0.64	0.87	0.88	0.76	0.78
sex	0.80	**0.88**	**0.88**	0.85	0.77	**0.88**	0.84	0.78	0.73
single	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.98	**0.99**
strains	0.92	0.92	0.92	**0.93**	0.89	0.92	0.92	**0.93**	0.87
support	0.90	0.90	0.90	0.81	0.71	0.90	**0.91**	0.83	0.79
surgery	**0.98**	0.94	0.95	0.95	**0.98**	**0.98**	**0.98**	0.95	0.93
transient	**0.99**	**0.99**	**0.99**	**0.99**	0.92	**0.99**	**0.99**	0.96	0.84
transport	**0.93**	**0.93**	**0.93**	0.92	0.92	**0.93**	**0.93**	0.90	0.92
ultrasound	0.84	0.84	**0.90**	0.81	0.67	0.85	0.87	0.84	0.71
variation	0.80	**0.88**	0.87	0.79	0.74	0.86	0.86	0.77	0.74
weight	0.47	**0.74**	0.68	0.63	0.66	0.71	**0.74**	0.67	0.66
white	0.49	**0.78**	0.70	0.65	0.63	0.74	0.70	0.63	0.71
Overall Accuracy	0.78	**0.85**	0.84	0.79	0.74	**0.85**	0.84	0.81	0.77

Table 4.4: P-values using the Pairwise T-test for Frequency Cutoff Results

		Naive Bayes				SVM			
		No Cutoff	2	4	6	No Cutoff	2	4	6
	Baseline	0.00002	0.00005	0.28568	0.03203	0.000001	0.00010	0.01031	0.35911
Naive Bayes	No Cutoff		0.01762	0.000001	0.000001	0.45972	0.00630	0.000001	0.000001
	2			0.000001	0.000001	0.01307	0.27443	0.000001	0.000001
	4				0.00003	0.000001	0.000001	0.01392	0.04722
	6					0.000001	0.000001	0.00002	0.00936
SVM	No Cutoff						0.00198	0.000001	0.000001
	2							0.000001	0.000001
	4								0.00033

Table 4.5: Analysis of Features

Cutoff	0	2	4	6
Avg. # Features in Feature Set	2455.48	1426.98	865.89	559.25
Avg. # Non-Zero Elements	50.77	44.92	38.19	32.77
Naive Bayes Accuracy	0.85	0.84	0.79	0.74
SVM Accuracy	0.85	0.84	0.81	0.77

To explore this, the target word *pathology* is analyzed. For fold two, the overall disambiguation accuracy is 100% when not using a cutoff and then decreases by 60% when using a cutoff of two. The training data contains 90 instances where 75 are classified as Pathology [C0030664], 14 classified as Pathology <3> [C0677042] and one classified as None. The test data contains ten instances each classified as Pathology [C0030664]. There exists approximately 5.1 features that exist in the training data only once and in the test instances. Table 4.6 shows the features that exist in the training data once and in the test instances - the starred instances are those that were misclassified after removing the features that only occur once from the feature set. These results show that the supervised learning model is using information from features that only occur once in the feature set.

Overall the results show that, regardless of the algorithm, not using a cutoff (or a cutoff of zero) returns a higher disambiguation accuracy, although using a frequency cutoff of two only reduces the accuracy by one percentage point it indicates that the CUIs that only occur once in the training data are affecting the accuracy of the supervised learning model.

Table 4.6: Features that occur once in the Training Data and in the Test Data

Instance 9307917*	**Instance 9421796**
Perfusion	Insight, NOS
Juxta-posed	Uterine Diseases
Percent Gradient	Blood Supply <2>
Imbrication	Separated From Cohabitee
Parameter	Killer Cells Natural
Perfusion, NEC	Uterus
Peak	Separated <2>
Rapid	Dilation Pathologic, NOS
	Uterine
	Lymphocytes
	Uterus, NEC
Instance 9376972	**Instance 9477404***
Vacuolar Myelopathy	Stress
Aids Patient	Preoperative
Spinal	Paper
Cytomegalovirus	Microsurgery
Worse	Very Low
Instance 9627012	**Instance 9698252***
Segment	Angigens CD95
Absences	Endocrine Cell, NOS
Foetal	Hormones
Placento	Immune
Herpes Virus 4 Human	Fetal Alcohol Syndrome
Instance 9743321	**Instance 9451461**
Non-specific	Closed <2>
Preliminary	Abscess
Closed Approach	Encephalitis
Instance 9547334*	**Instance 9621635**
Attenuation	Critic
Low Frequency	Critic, NOS
Colon, NEC	
Colon	

4.2.2 MMI Cutoff Results

This section discusses the results of the MMI cutoff experiment. The hypothesis of this experiment is that using a MMI cutoff will reduce the amount of noise in the feature set increasing the overall disambiguation accuracy. An MMI score quantifies how relevant a CUIs is in describing the overall topic of an instance. The assumption is that CUIs that are more relevant are better indicators of describing its context and therefore better indicators of the context in which a target word is used.

To test this hypothesis, this section analyzes the results of using a MMI cutoff of 10 and 20, and compares them to the results of not using a cutoff. Table 4.7 shows majority sense baseline and the results for these experiments using the SVM and Naive Bayes algorithm. Table 4.8 shows the statistical significance between the results.

The results show that using a MMI cutoff of 10 and 20 obtains an accuracy of 84% and 82% respectively regardless of the algorithm. This is a statistically significant higher disambiguation accuracy than the baseline.

Further analysis of the data shows that the average number of features in the feature set and the average number of non-zero elements in the test vectors are greatly reduced when using the MMI cutoff. Table 4.9 shows the average number of features in the feature set, the average number of non-zero elements in the test vectors when using each of the MMI cutoffs and no cutoff. The results show that when using a MMI cutoff of 10, the feature test contains 1489.92 which is almost 60% less features than not using the cutoff and the accuracy only decreases by one percentage point. The average number of non-zero elements in the test vectors is 18.583 which means that with almost 70% less features K-CUI is able classify each of the test vectors with only a one percentage point decrease in overall accuracy.

The overall conclusion of the MMI experiments is that using MMI cutoff significantly reduces the noise in the feature set while maintaining the overall accuracy of the system.

4.2.3 Semantic Similarity Cutoff Results

This section discusses the results of the similarity cutoff experiment. The hypothesis of this experiment is that using a semantic similarity cutoff will reduce the amount of noise in the feature set increasing the overall disambiguation accuracy. Semantic similarity

Table 4.7: MMI Cutoff Results

		Naive Bayes			SVM		
	Baseline	No Cutoff	MMI		No Cutoff	MMI	
adjustment	0.62	0.65	**0.71**	0.68	0.65	0.66	0.59
association	**1.00**	0.98	0.99	**1.00**	**1.00**	**1.00**	**1.00**
blood pressure	**0.54**	**0.54**	**0.54**	0.51	0.51	0.53	0.47
cold	0.86	**0.89**	**0.89**	0.86	0.88	0.88	0.88
condition	0.90	0.90	**0.92**	**0.92**	0.90	0.90	0.90
culture	0.89	**0.91**	**0.91**	0.86	0.89	**0.91**	0.90
degree	0.63	0.81	0.74	0.70	**0.82**	0.74	0.70
depression	**0.85**	0.83	**0.85**	0.76	0.84	**0.85**	0.84
determination	0.79	0.79	**0.87**	0.79	0.84	0.82	0.76
discharge	0.74	**0.93**	0.90	0.88	0.92	0.92	0.85
energy	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.50	0.69	0.67	0.58	**0.74**	0.65	0.62
extraction	0.82	0.84	**0.85**	0.83	0.84	0.83	0.83
failure	**0.71**	0.68	0.68	0.67	**0.71**	0.68	0.70
fat	0.71	0.74	0.81	0.78	0.78	0.74	**0.82**
fit	0.82	0.85	0.87	**0.89**	0.86	0.86	0.85
fluid	**1.00**	0.98	0.99	**1.00**	**1.00**	**1.00**	**1.00**
frequency	0.94	0.94	0.93	0.94	**0.95**	0.94	0.93
ganglion	0.93	0.94	**0.97**	0.90	0.93	0.91	0.93
glucose	**0.91**	0.90	0.90	0.87	**0.91**	0.90	0.90
growth	0.63	**0.72**	0.64	0.53	0.70	0.65	0.61
immunosuppression	0.59	0.80	0.77	0.69	**0.87**	0.77	0.62
implantation	0.81	**0.94**	**0.94**	0.87	0.92	0.90	0.88
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
japanese	0.73	**0.78**	**0.78**	**0.78**	**0.78**	0.75	0.77
lead	0.71	0.90	0.84	0.87	0.93	**0.95**	0.92
man	0.58	0.83	0.76	0.74	**0.84**	0.83	0.70
mole	0.83	0.84	0.84	0.85	0.86	0.85	**0.91**
mosaic	0.52	0.78	0.74	0.74	0.76	**0.82**	0.70
nutrition	0.45	**0.47**	0.38	0.42	0.42	0.43	0.46
pathology	0.85	0.85	**0.88**	0.83	0.83	0.85	0.83
pressure	**0.96**	0.95	0.95	0.88	**0.96**	0.94	**0.96**
radiation	0.61	0.84	0.80	0.70	**0.86**	0.78	0.73
reduction	0.89	0.89	**0.90**	**0.90**	0.89	**0.90**	**0.90**
repair	0.52	**0.91**	0.84	0.79	0.84	0.77	0.70
resistance	**0.97**	0.96	0.95	**0.97**	**0.97**	**0.97**	**0.97**
scale	0.65	0.77	**0.78**	0.73	0.75	0.63	0.72
secretion	**0.99**	**0.99**	**0.99**	0.95	**0.99**	**0.99**	**0.99**
sensitivity	0.49	**0.92**	0.87	0.82	0.87	0.81	0.75
sex	0.80	**0.88**	0.85	0.82	**0.88**	0.83	0.83
single	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
strains	**0.92**	**0.92**	0.91	**0.92**	**0.92**	**0.92**	**0.92**
support	**0.90**	**0.90**	0.89	0.85	**0.90**	**0.90**	0.87
surgery	**0.98**	0.94	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
transient	**0.99**	**0.99**	0.95	**0.99**	**0.99**	**0.99**	**0.99**
transport	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	0.92
ultrasound	0.84	0.84	0.82	0.75	0.85	0.84	**0.86**
variation	0.80	**0.88**	0.87	0.86	0.86	0.86	0.87
weight	0.47	0.74	0.74	0.71	0.71	**0.76**	0.64
white	0.49	**0.78**	0.70	0.76	0.74	0.73	0.76
Overall Accuracy	0.78	**0.85**	0.84	0.82	**0.85**	0.84	0.82

Table 4.8: P-values using the Pairwise T-test for MMI Results

		Naive Bayes		SVM	
		MMI = 10	MMI = 20	MMI = 10	MMI = 20
Naive Bayes	MMI = 10				
	MMI = 20	0.00002			
	None	0.04904	0.00001		
SVM	MMI = 10	0.21231	0.00067		
	MMI = 20	0.00318	0.27690	0.00937	
	None	0.02866	0.00001	0.00310	0.00019
baseline		0.00001	0.00266	0.00004	0.00011

Table 4.9: Average Number of Features and Non-Zero Elements in Test Vectors

	MMI Cutoff = 10	MMI Cutoff = 20	No Cutoff
Average # Features	1489.92	669.30	3752.64
Average # Non-Zero Elements	18.58	5.91	63.49
Overall Accuracy	0.84	0.82	0.85

measures assign a score to how similar two concepts are to each other. The assumption behind using a semantic similarity cutoff is that CUIs that have a high similarity score to one of the possible concepts of the target word will be a good indicator of the target words concept.

A-CUI uses the following semantic similarity measures from the UMLS::Similarity package in these experiments: a simple path measure and the measure proposed by [Wu and Palmer, 1994]. The path measure is the reciprocal of the number of nodes between two concepts. The similarity measure proposed by Wu & Palmer (WUP) is twice the depth of the two concepts least common subsumer (LCS) divided by the product of the depths of the individual concepts. The LCS is the most specific concept two concepts share as an ancestor. The UMLS::Similarity package uses the PAR/CHD relations in SNOMED-CT to obtain the path information for the semantic similarity measures because it is the largest source available in the UMLS. Each of these measures return a semantic similarity score between zero and one where one indicates the two concepts are synonymous. This experiments investigates using a cutoff score of 0.1 and 0.2.

Table 4.10: Semantic Similarity Cutoff Results

		Naive Bayes					SVM				
	Baseline	No	Path		WUP		No	Path		WUP	
		cutoff	0.1	0.2	0.1	0.2	cutoff	0.1	0.2	0.1	0.2
adjustment	0.62	0.65	0.64	0.59	**0.67**	0.64	0.65	0.64	0.66	**0.67**	0.60
association	**1.00**	0.98	0.97	0.96	0.98	0.96	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
blood pressure	0.54	0.54	0.54	0.44	**0.55**	0.35	0.51	0.51	0.43	0.52	0.40
cold	0.86	**0.89**	0.88	0.68	0.87	0.87	0.88	0.88	0.87	0.88	0.88
condition	**0.90**	**0.90**	0.89	0.88	**0.90**	**0.90**	**0.90**	0.89	0.89	0.89	0.87
culture	0.89	0.91	0.91	0.89	**0.92**	0.88	0.89	0.90	0.89	0.89	0.88
degree	0.63	0.81	0.74	0.63	0.74	0.73	**0.82**	0.76	0.58	0.72	0.68
depression	**0.85**	0.83	0.75	0.76	0.78	0.77	0.84	0.80	0.76	0.82	0.83
determination	0.79	0.79	0.79	0.79	0.79	0.79	**0.84**	0.79	0.79	0.79	0.79
discharge	0.74	**0.93**	0.91	0.69	**0.93**	0.89	0.92	0.87	0.74	0.89	0.84
energy	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.50	0.69	0.65	0.56	0.68	0.66	**0.74**	0.69	0.57	0.70	0.68
extraction	0.82	**0.84**	0.83	0.77	0.83	0.83	**0.84**	0.80	0.82	0.81	0.81
failure	0.71	0.68	0.75	0.71	0.74	0.77	0.71	0.73	0.71	0.74	**0.78**
fat	0.71	0.74	0.69	0.69	0.76	0.70	**0.78**	0.71	0.71	0.74	0.73
fit	0.82	0.85	0.82	0.81	0.85	0.87	0.86	**0.89**	0.82	0.85	0.85
fluid	**1.00**	0.98	0.98	**1.00**	0.98	0.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
frequency	0.94	0.94	0.94	0.94	0.94	0.94	**0.95**	**0.95**	0.94	**0.95**	**0.95**
ganglion	0.93	0.94	0.91	0.93	0.93	**0.95**	0.93	0.94	0.93	0.93	**0.95**
glucose	0.91	0.90	0.90	0.91	0.90	0.87	0.91	**0.92**	0.91	0.91	0.91
growth	0.63	**0.72**	0.63	0.63	0.63	0.63	0.70	0.63	0.63	0.63	0.63
immunosuppression	0.59	0.80	0.67	0.60	0.70	0.71	**0.87**	0.66	0.60	0.77	0.69
implantation	0.81	**0.94**	0.88	0.81	0.93	0.90	0.92	0.91	0.78	**0.94**	0.93
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
japanese	0.73	**0.78**	0.71	0.74	0.77	0.73	**0.78**	0.72	0.74	0.77	**0.78**
lead	0.71	0.90	0.78	0.81	0.84	0.83	**0.93**	0.82	0.81	0.86	0.88
man	0.58	0.83	0.82	0.73	0.81	0.82	**0.84**	0.83	0.72	0.83	0.83
mole	0.83	0.84	0.81	0.82	0.82	0.83	**0.86**	0.80	0.83	0.84	**0.86**
mosaic	0.52	**0.78**	0.72	0.63	0.74	0.70	0.76	0.70	0.67	0.69	0.69
nutrition	0.45	**0.47**	0.38	0.35	0.39	0.36	0.42	0.32	0.35	0.34	0.37
pathology	**0.85**	**0.85**	0.81	0.81	0.81	0.79	0.83	0.80	0.83	0.78	0.79
pressure	**0.96**	0.95	0.74	0.68	0.75	0.65	**0.96**	0.74	0.70	0.74	0.68
radiation	0.61	0.84	0.80	0.52	0.80	0.70	**0.86**	0.78	0.59	0.77	0.68
reduction	0.89	0.89	0.91	0.91	0.89	0.91	0.89	0.90	**0.92**	0.89	0.89
repair	0.52	**0.91**	0.83	0.50	0.87	0.84	0.84	0.77	0.57	0.83	0.81
resistance	**0.97**	0.96	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
scale	0.65	0.77	0.73	0.66	0.78	0.74	0.75	0.66	0.71	**0.79**	0.72
secretion	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
sensitivity	0.49	**0.92**	0.67	0.52	0.91	0.66	0.87	0.70	0.47	0.84	0.60
sex	0.80	0.88	0.86	0.80	0.86	0.84	0.88	0.89	0.81	**0.90**	0.86
single	**0.99**	0.98	0.98	0.98	0.98	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
strains	**0.92**	**0.92**	0.91	**0.92**	**0.92**	**0.92**	**0.92**	0.91	**0.92**	**0.92**	**0.92**
support	0.90	0.90	0.91	**0.93**	0.91	0.89	0.90	0.90	**0.93**	0.90	0.90
surgery	**0.98**	0.94	0.93	**0.98**	0.93	0.93	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
transient	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
transport	**0.93**	**0.93**	**0.93**	0.91	**0.93**	0.92	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
ultrasound	0.84	0.84	0.87	0.78	0.82	0.85	0.85	**0.89**	0.82	0.86	0.88
variation	0.80	**0.88**	0.81	0.78	0.86	0.80	0.86	0.76	0.81	0.79	0.77
weight	0.47	**0.74**	0.47	0.47	0.47	0.47	0.71	0.47	0.47	0.47	0.47
white	0.49	**0.78**	0.63	0.64	0.63	0.64	0.74	0.69	0.61	0.73	0.73
Overall Accuracy	0.78	**0.85**	0.81	0.77	0.83	0.81	**0.85**	0.81	0.78	0.83	0.81

Table 4.11: P-values using the Pairwise T-test for Semantic Similarity Cutoff Results

			Naive Bayes (NB)					SVM				
			No	Path		WUP		No	Path		WUP	
			cutoff	0.1	0.2	0.1	0.2	cutoff	0.1	0.2	0.1	0.2
	baseline		.00002	.00898	.12592	.00131	.03668	.000001	.00482	.40020	.00104	.00999
NB	No	Cutoff		.00002	.000001	.00114	.00002	.40454	.00013	.00001	.00185	.00035
	Path	0.1			.00018	.00210	.15678	.000001	.22413	.00212	.00284	.46972
	Path	0.2				.00002	.00048	.000001	.00002	.00587	.000001	.00001
	WUP	0.1					.00160	.00092	.02488	.00032	.48207	.04751
	WUP	0.2						.000001	.06600	.00736	.00035	.08387
SVM	No	Cutoff							.00002	.000001	.00048	.00008
	Path	0.1								.00063	.00584	.31842
	Path	0.2									.00010	.00035
	WUP	0.1										.00825

The similarity cutoff is evaluated using both the Naive Bayes and the SVM algorithms and compared with the majority sense baseline and the results obtained when no cutoff is used. Table 4.10 shows the majority sense baseline and results for these experiments using the Naive Bayes and the SVM algorithm. Table 4.11 shows the statistical significance between the results.

The results show that using a similarity cutoff of 0.1 and 0.2 using the path measure (path) obtains an accuracy of 81% and 77% respectively for the Naive Bayes and 81% and 78% for the SVM. When using the measure proposed by [Wu and Palmer, 1994] the accuracy increases. Both the Naive Bayes and SVM results report an accuracy of 83% and 81% using the similarity cutoff of 0.1 and 0.2 respectively. There is a statistically significant difference between the baseline and the results obtained using similarity cutoff except when using the path based measure with a similarity cutoff of 0.1.

The similarity cutoff of 0.1 returns a higher disambiguation accuracy than using a cutoff of 0.2 for both semantic similarity measures, and the semantic similarity measure proposed by [Wu and Palmer, 1994] returns a higher disambiguation accuracy than using the path based measure regardless of the algorithm. The differences in accuracy for both these results is statistically significant.

The best semantic similarity results use the [Wu and Palmer, 1994] measure with a

cutoff of 0.1. The overall results show an 83% accuracy for both the Naive Bayes and SVM algorithms which is three percentage points lower than using no cutoff at all.

Table 4.12: Average Number of Features and Non-Zero Elements in Test Vectors

	Path		WUP		No Cutoff
	0.1	0.2	0.1	0.2	
Average # Features	621.22	38.85	1026.38	498.65	3752.64
Average # Non-Zero Elements	18.37	1.70	25.62	13.27	63.49
Overall Accuracy	0.82	0.79	0.83	0.82	0.85

Further analysis of the data shows that the average number of features in the feature set and the average number of non-zero elements in the test vectors are greatly reduced when using the semantic similarity cutoff. Table 4.12 shows the average number of features in the feature set, the average number of non-zero elements in the test vectors when using each of the similarity cutoffs and no cutoff. The results show that when using the measure proposed by [Wu and Palmer, 1994] with a cutoff of 0.1, the feature test contains 1026.38 which is almost 70% less features than not using the cutoff and the accuracy only decreases by two percentage point. The average number of non-zero elements in the test vectors is 25.62 which means that with almost 60% less features K-CUI is able classify each of the test vectors with only a two percentage point decrease in overall accuracy.

The semantic similarity score is obtained using the UMLS::Similarity package. A disadvantage of the package is that the semantic similarity can not be taken between all of the CUIs in the UMLS. The similarity can only be taken between CUIs in a pre-specified subset of the UMLS. These experiments use the concepts from SNOMED-CT which is the largest single source in the UMLS. Unfortunately, not all of the possible concepts for a given target word in the NLM-WSD dataset exist in SNOMED-CT. Out of the 50 target words, only 23 of the target words had all possible concepts in SNOMED-CT, 20 had at least one of their possible concepts in SNOMED-CT and five did not have any. Table 4.13 shows the list of target words and the possible concepts that do not exist in the SNOMED-CT. The five target words that did not have any of their possible concept in SNOMED-CT are:

- determination

- growth
- resistance
- secretion
- surgery

Even though not all of the possible concepts have an associated CUI in SNOMED-CT, the highest overall accuracy reported for the semantic similarity results is 83% which is only two percentage points lower than not using a cutoff and five percentage points higher than the baseline. This indicates that there existed enough information in the possible concepts that do exist in SNOMED-CT to disambiguate between them. The results have the potential to increase if the semantic similarity could be taken between all of a target words possible concepts.

The overall conclusion of the semantic similarity cutoff experiments is this type of cutoff significantly reduces the noise in the feature set while maintaining the overall accuracy of the system.

4.3 Comparison with General English Features

This section discusses the comparison between using the general English feature unigrams and the biomedical feature CUIs. Unigrams consist of word level information, the feature set contains highly frequent content words that surround the target word. A stoplist is used to determine which words are content words - Appendix F shows the list of stopwords used for this experiment.

CUIs provide unambiguous term level information and contain less noise than using the individual words. CUIs encompass not just a single word but a term. Noise is reduced because each CUI contains some content information from the biomedical domain. Non-content words such as determiners (*the*, *a*, *an*) and prepositions (*to*, *for*, *in*) do not have an associated CUI in the UMLS as well as content words that are not biomedical in nature. The CUIs are obtained using MetaMap which does not attempt to disambiguate terms that map to more than one CUI. K-CUI uses all of the possible mappings provided by MetaMap in the feature set.

The purpose of this experiment is to compare the performance of CUIs and unigrams by conducting the following experiments:

Table 4.13: NLM-WSD Concepts Not in SNOMED-CT

Target word	CUI
adjustment	Psychological adjustment [C0683269]
association	Relationship by association [C0699792]
blood pressure	Arterial pressure [C0428878]
culture	Anthropological Culture [C0010453]
degree	Degree [C0542560]
determination	Adjudication [C0680730] Determination [C0243075]
energy	Energy (physics [C0542479]
extraction	Extraction [C0684295]
failure	Failure [C0699796]
ganglion	Ganglia [C0017067]
growth	Growth 1 [C0018270] Growth 2 [C0220844]
inhibition	Psychological inhibition [C0021467]
japanese	Japanese Population [C0022342]
lead	Lead measurement, quantitative [C0373667]
mole	Mole the mammal [C0026386] Benign melanocytic nevus of skin [C0349514]
mosaic	Mosaicism [C0026578] Mosaic [C0700058]
nutrition	Science of nutrition [C0028707]
pathology	Pathology [C0677042]
radiation	Radiation therapy [C0034618]
resistance	Resistance 1 [C0683598] resistance 2 [C0237834]
secretion	Bodily Secretions [C0036537] Secretion [C0687157]
sensitivity	Statistical Sensitivity [C0036667]
sex	Coitus [C0036862]
single	Unmarried [C0087136]
surgery	Surgery specialty [C0038894] Surgery [C0600001]
transient	Transient Population Group [C0040704]
ultrasound	Ultrasonic Shockwave [C0041621]

Table 4.14: Comparison between CUI and Unigram Results

	CUIs						Unigrams					
	Naive Bayes			SVM			Naive Bayes			SVM		
	0	2	4	0	2	4	0	2	4	0	2	4
adjustment	0.65	0.65	0.49	0.65	0.67	0.60	0.69	**0.70**	0.60	0.69	0.69	0.60
association	0.98	0.99	**1.00**	**1.00**	**1.00**	**1.00**	0.97	0.99	**1.00**	**1.00**	**1.00**	**1.00**
blood pressure	**0.54**	0.45	0.51	0.51	0.40	0.50	0.51	0.52	0.41	0.47	0.46	0.45
cold	0.89	**0.90**	0.86	0.88	0.88	0.88	0.88	0.89	**0.90**	0.88	0.88	0.89
condition	**0.90**	0.89	0.88	**0.90**	0.89	0.89	**0.90**	**0.90**	0.87	**0.90**	0.87	0.86
culture	0.91	0.92	0.88	0.89	0.90	0.88	0.89	0.91	0.91	0.89	0.92	**0.94**
degree	0.81	0.81	0.66	0.82	0.78	0.73	0.80	0.78	0.75	0.80	**0.83**	0.73
depression	0.83	0.81	**0.87**	0.84	0.84	0.83	0.77	0.84	0.85	0.84	0.86	0.85
determination	0.79	0.82	0.53	**0.84**	0.76	0.74	0.81	0.68	0.68	0.80	0.82	0.79
discharge	0.93	0.94	**0.95**	0.92	0.94	0.86	0.94	0.91	0.89	0.90	0.91	0.87
energy	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.69	0.65	0.64	0.74	0.63	0.68	**0.77**	0.68	0.61	0.75	0.64	0.69
extraction	0.84	**0.86**	0.84	0.84	0.83	0.80	**0.86**	**0.86**	0.82	0.84	0.84	0.77
failure	0.68	0.68	0.58	**0.71**	0.70	0.67	0.70	0.60	0.43	0.69	0.65	0.67
fat	0.74	0.73	0.67	0.78	0.76	0.73	0.81	0.75	0.67	**0.82**	0.73	0.67
fit	0.85	0.86	0.70	0.86	0.85	0.82	**0.88**	0.87	0.72	0.85	0.85	0.74
fluid	0.98	0.99	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
frequency	0.94	**0.95**	0.76	**0.95**	**0.95**	**0.95**	0.93	**0.95**	0.84	0.94	**0.95**	0.92
ganglion	0.94	**0.97**	0.92	0.93	0.95	0.96	0.94	0.95	0.93	0.94	0.95	0.95
glucose	0.90	0.90	0.87	**0.91**	0.88	0.84	0.90	**0.91**	0.90	**0.91**	0.87	0.76
growth	0.72	0.73	0.65	0.70	0.66	0.63	**0.74**	0.68	0.73	0.67	0.66	0.70
immunosuppression	0.80	0.79	0.77	**0.87**	0.76	0.76	0.73	0.71	0.74	0.77	0.65	0.65
implantation	**0.94**	0.91	0.79	0.92	0.92	0.91	0.92	0.93	0.57	0.89	0.93	0.82
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
japanese	**0.78**	0.75	0.64	**0.78**	0.73	0.75	**0.78**	0.76	0.67	0.77	0.73	0.56
lead	0.90	0.83	0.74	**0.93**	0.92	0.86	0.85	0.72	0.45	0.85	0.75	0.29
man	0.83	0.80	0.72	**0.84**	0.83	0.75	0.79	0.74	0.46	0.81	0.82	0.58
mole	0.84	0.85	0.69	0.86	0.85	0.81	0.84	0.84	0.85	0.84	**0.89**	0.84
mosaic	0.78	0.77	0.69	0.76	0.75	0.75	0.78	0.75	0.63	**0.85**	0.82	0.76
nutrition	0.47	0.42	0.34	0.42	0.38	0.42	**0.50**	0.38	0.29	0.47	0.38	0.35
pathology	0.85	0.73	0.59	0.83	0.84	0.72	0.84	0.81	0.74	**0.86**	**0.86**	0.73
pressure	0.95	**0.96**	**0.96**	**0.96**	**0.96**	0.89	0.95	**0.96**	0.68	0.72	**0.96**	**0.96**
radiation	0.84	0.82	0.75	**0.86**	0.84	0.71	0.82	0.74	0.69	0.83	0.78	0.63
reduction	0.89	0.89	0.86	0.89	0.89	0.83	0.89	0.89	**0.92**	0.89	0.89	0.90
repair	**0.91**	0.88	0.84	0.84	0.83	0.80	0.83	0.87	0.79	0.81	0.86	0.67
resistance	0.96	**0.97**	**0.97**	**0.97**	**0.97**	0.96	0.96	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
scale	0.77	0.77	0.79	0.75	0.74	0.73	**0.81**	0.80	0.70	0.78	0.79	0.66
secretion	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
sensitivity	**0.92**	0.88	0.83	0.87	0.88	0.76	0.90	0.90	0.77	0.90	0.87	0.79
sex	0.88	0.88	0.85	0.88	0.84	0.78	0.83	0.88	0.82	**0.89**	0.85	0.88
single	0.98	**0.99**	**0.99**	**0.99**	**0.99**	0.98	0.98	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
strains	0.92	0.92	**0.93**	0.92	0.92	**0.93**	0.92	0.92	0.92	0.92	0.92	**0.93**
support	0.90	0.90	0.81	0.90	**0.91**	0.83	0.90	0.90	0.80	0.90	0.86	0.75
surgery	0.94	0.95	0.95	**0.98**	**0.98**	0.95	0.94	0.96	0.96	**0.98**	**0.98**	**0.98**
transient	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.96	0.97	**0.99**	**0.99**	**0.99**	**0.99**	0.98
transport	**0.93**	**0.93**	0.92	**0.93**	**0.93**	0.90	**0.93**	**0.93**	0.91	**0.93**	0.92	0.87
ultrasound	0.84	**0.90**	0.81	0.85	0.87	0.84	0.81	0.82	0.73	0.86	0.82	0.78
variation	0.88	0.87	0.79	0.86	0.86	0.77	0.90	**0.96**	0.80	0.84	0.89	0.82
weight	0.74	0.68	0.63	0.71	0.74	0.67	**0.79**	0.73	0.67	**0.79**	0.78	0.54
white	**0.78**	0.70	0.65	0.74	0.70	0.63	0.68	0.69	0.59	0.72	0.62	0.57
Overall Accuracy	**0.85**	0.84	0.79	**0.85**	0.84	0.81	**0.85**	0.84	0.77	**0.85**	0.84	0.78

- Feature set: CUIs
  - Frequency Cutoff: 0, 2, 4
  - Algorithms: Naive Bayes and SVM
- Feature set: Unigrams
  - Frequency Cutoff: 0, 2, 4
  - Algorithms: Naive Bayes and SVM
- Feature set: CUIs + Unigrams
  - No Frequency Cutoff
  - Algorithms: Naive Bayes and SVM

Table 4.14 shows the results of the experiments using the individual features sets and the majority sense baseline, and Table 4.15 shows the statistical significance between the results. Table 4.16 shows the results for the feature set containing both the CUIs and the unigrams and the individual feature sets using no frequency cutoff.

Table 4.15: P-Values using the Pairwise T-test for CUI and Unigram Results

		Unigram Results					
		Naive Bayes			SVM		
CUI Results	Cutoff	0	2	4	0	2	4
Naive Bayes	0	0.20338	0.00746	0.000001	0.16554	0.01280	0.00001
	2	0.16220	0.15437	0.000001	0.32513	0.18191	0.00001
	4	0.000001	0.000001	0.08532	0.00001	0.00007	0.27209
SVM	0	0.18627	0.01612	0.000001	0.13144	0.01625	0.00001
	2	0.09722	0.31001	0.00001	0.18910	0.31811	0.00016
	4	0.000001	0.00075	0.00366	0.000001	0.00043	0.01678

The results show that both the unigram and the CUI results return an overall accuracy of 85% when using both the Naive Bayes and SVM algorithms. As the frequency cutoff increases though the unigram results decrease at a much faster rate than the CUI results. Neither feature set benefits from using a cutoff, the more features available the higher the disambiguation accuracy of the method.

The results of using the unigrams and CUIs in a single feature set show an increase in overall accuracy of one percentage point when using both the Naive Bayes and SVM algorithms indicating the unigrams and CUIs are each providing information that is used in the disambiguation process.

Table 4.16: Combination CUI + Unigram Results

	Naive Bayes			SVM		
	CUIs	Unigrams	CUIs+Unigrams	CUIs	Unigrams	CUIs+Unigrams
adjustment	0.65	0.69	0.68	0.65	0.69	**0.71**
association	0.98	0.97	0.97	**1.00**	**1.00**	**1.00**
blood pressure	**0.54**	0.51	0.53	0.51	0.47	0.47
cold	**0.89**	0.88	0.88	0.88	0.88	0.88
condition	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**
culture	**0.91**	0.89	0.90	0.89	0.89	0.89
degree	0.81	0.80	**0.83**	0.82	0.80	**0.83**
depression	0.83	0.77	0.77	**0.84**	**0.84**	**0.84**
determination	0.79	0.81	0.82	**0.84**	0.80	0.82
discharge	0.93	0.94	**0.95**	0.92	0.90	0.92
energy	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.69	0.77	**0.78**	0.74	0.75	0.74
extraction	0.84	**0.86**	**0.86**	0.84	0.84	0.84
failure	0.68	0.70	0.71	0.71	0.69	**0.72**
fat	0.74	0.81	0.79	0.78	**0.82**	0.81
fit	0.85	**0.88**	0.85	0.86	0.85	0.84
fluid	0.98	0.99	0.99	**1.00**	**1.00**	**1.00**
frequency	0.94	0.93	0.93	**0.95**	0.94	0.94
ganglion	**0.94**	**0.94**	**0.94**	0.93	**0.94**	**0.94**
glucose	0.90	0.90	0.90	**0.91**	**0.91**	**0.91**
growth	0.72	**0.74**	**0.74**	0.70	0.67	0.72
immunosuppression	0.80	0.73	0.79	**0.87**	0.77	**0.87**
implantation	**0.94**	0.92	**0.94**	0.92	0.89	0.91
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
japanese	0.78	0.78	**0.79**	0.78	0.77	0.78
lead	0.90	0.85	0.87	**0.93**	0.85	0.88
man	0.83	0.79	0.83	**0.84**	0.81	0.82
mole	0.84	0.84	**0.86**	**0.86**	0.84	0.85
mosaic	0.78	0.78	0.78	0.76	**0.85**	0.83
nutrition	0.47	**0.50**	0.48	0.42	0.47	0.45
pathology	0.85	0.84	0.84	0.83	**0.86**	0.85
pressure	0.95	0.95	0.95	**0.96**	0.72	**0.96**
radiation	0.84	0.82	0.82	**0.86**	0.83	**0.86**
reduction	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**
repair	**0.91**	0.83	0.89	0.84	0.81	0.82
resistance	0.96	0.96	0.96	**0.97**	**0.97**	**0.97**
scale	0.77	**0.81**	0.80	0.75	0.78	0.79
secretion	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
sensitivity	0.92	0.90	**0.93**	0.87	0.90	0.92
sex	0.88	0.83	0.86	0.88	0.89	**0.91**
single	0.98	0.98	0.98	**0.99**	**0.99**	**0.99**
strains	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**
support	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**
surgery	0.94	0.94	0.93	**0.98**	**0.98**	**0.98**
transient	**0.99**	0.97	0.97	**0.99**	**0.99**	**0.99**
transport	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
ultrasound	0.84	0.81	0.83	0.85	0.86	**0.87**
variation	0.88	0.90	**0.91**	0.86	0.84	0.84
weight	0.74	0.79	**0.80**	0.71	0.79	0.78
white	**0.78**	0.68	0.72	0.74	0.72	0.71
Overall Accuracy	0.85	0.85	**0.86**	0.85	0.85	**0.86**

4.4 Comparison with Related Work

There has been previous work evaluated on various subsets of the NLM-WSD dataset. [Leroy and Rindflesch, 2004] evaluate their method using those target words in the dataset that have a majority sense of 65% or less, which consists of 15 out of the 50 words and is referred to as the Leroy subset. [Liu et al., 2004] evaluate their method using 22 out of the 50 words in the dataset; referred to as the Liu subset. [Joshi et al., 2005] evaluate their approach using both the Liu and Leroy subset. This union is referred to as the Joshi subset. [Savova et al., 2008] evaluate their method using 42 out of the 50 target words in the dataset; referred to as the Savova subset. [Stevenson et al., 2008] evaluate theirs using the entire NLM-WSD dataset.

Table 4.17 shows the results of methods proposed by the previous work and those obtained using K-CUI. [Leroy and Rindflesch, 2004], [Joshi et al., 2005] and [Stevenson et al., 2008] are directly comparable with each other and K-CUI using their respective subsets. The results reported by [Liu et al., 2004] and [Savova et al., 2008] are not directly comparable because the results reported are the highest per word accuracy over all feature sets and algorithms making it difficult to determine which method and feature set returned the highest accuracy overall.

There are four different K-CUI results reported in order to compare the feature sets directly with the related work.

- the feature set containing the CUIs in the same abstract as the target word with no cutoff using Naive Bayes
- the feature set containing the CUIs in the sentence as the target word with no cutoff using Naive Bayes
- the feature set containing the CUIs in the same abstract as the target word with no cutoff using SVMs
- the feature set containing the CUIs in the sentence as the target word with no cutoff using SVMs

The researchers discussed in this section evaluate their work using the 10-fold cross validation option available in the WEKA data mining package. In this evaluation, the feature set is created using the entire subset of the NLM-WSD dataset being used and then the data is split into the blocks for training and testing purposes. The problem

Table 4.17: K-CUI and Related Work Results

	Base-line	K-CUI Naive Bayes		K-CUI SVMs		Stevenson et. al.,	Savova et. al.,	Joshi et. al.,	Leroy and Rindflesch,	Liu et. al.,
		abst.	sent.	abst.	sent.	2008	2008	2005	2004	2004
adjustment	0.62	0.65	0.73	0.65	0.71	0.73	**0.75**	0.71	0.57	
association	**1.00**	0.98	**1.00**	**1.00**	**1.00**	**1.00**				
blood pressure	0.54	0.54	0.59	0.51	**0.62**	0.53	**0.62**	0.53	0.46	
cold	0.86	0.89	0.88	0.88	0.88	0.88	0.89	0.90		**0.91**
condition	0.90	0.90	**0.92**	0.90	0.91	0.89	0.91			
culture	0.89	0.91	0.88	0.89	0.89	**0.95**	0.94			
degree	0.63	0.81	0.83	0.82	0.82	0.93	0.96	0.89	0.68	**0.98**
depression	0.85	0.83	0.85	0.84	0.85	0.86	**0.90**	0.86		0.89
determination	0.79	0.79	0.74	0.84	0.78	**0.87**	**0.87**			
discharge	0.74	0.93	0.88	0.92	0.82	0.94	**0.95**	**0.95**		0.91
energy	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.98				
evaluation	0.50	0.69	0.60	0.74	0.71	**0.81**	0.77	0.69	0.57	
extraction	0.82	0.84	0.84	0.84	0.82	0.85	0.87	0.84		**0.89**
failure	0.71	0.68	0.67	0.71	0.72	0.73	**0.75**			
fat	0.71	0.74	0.80	0.78	0.83	0.84	0.83	0.84		**0.86**
fit	0.82	0.85	0.85	0.86	0.86	**0.88**	**0.88**			
fluid	**1.00**	0.98	**1.00**	**1.00**	**1.00**	**1**				
frequency	0.94	0.94	0.94	0.95	0.94	0.94	**0.96**			
ganglion	0.93	0.94	0.95	0.93	0.95	**0.96**	**0.96**			
glucose	**0.91**	0.90	0.90	**0.91**	0.90	**0.91**	**0.91**			
growth	0.63	0.72	0.70	0.70	0.67	0.72	0.72	0.71	0.62	**0.72**
immunosuppression	0.59	0.80	0.73	**0.87**	0.76	0.81	0.84	0.80	0.63	
implantation	0.81	0.94	0.87	0.92	0.85	0.91	**0.96**	0.94		0.90
inhibition	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**				
japanese	0.73	0.78	0.79	0.78	0.78	0.77	0.77	0.77		**0.79**
lead	0.71	0.90	0.85	0.93	0.82	**0.94**	0.93	0.89		0.91
man	0.58	0.83	0.88	0.84	0.86	0.86	0.87	0.89	0.80	**0.91**
mole	0.83	0.84	0.82	0.86	0.86	0.88	**0.96**	0.95		0.911
mosaic	0.52	0.78	0.68	0.76	0.78	0.85	0.87	0.87	0.66	**0.88**
nutrition	0.45	0.47	0.55	0.42	0.45	0.57	0.49	0.52	0.48	**0.58**
pathology	0.85	0.85	0.83	0.83	0.82	0.86	0.87	0.85		**0.88**
pressure	**0.96**	0.95	**0.96**	**0.96**	**0.96**	0.95	**0.96**			
radiation	0.61	0.84	0.75	**0.86**	0.66	0.85	0.82	0.82	0.72	
reduction	0.89	0.89	0.89	0.89	0.89	0.88	**0.91**	**0.91**		**0.91**
repair	0.52	**0.91**	0.73	0.84	0.72	0.86	0.89	0.87	0.81	0.76
resistance	**0.97**	0.96	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**			
scale	0.65	0.77	0.80	0.75	0.81	0.88	0.82	0.81	0.84	**0.91**
secretion	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**				
sensitivity	0.49	0.92	0.76	0.87	0.72	**0.93**	0.92	0.88	0.70	
sex	0.80	0.88	0.79	0.88	0.84	0.87	**0.91**	0.88		0.89
single	**0.99**	0.98	**0.99**	**0.99**	**0.99**	**0.99**				
strains	0.92	0.92	0.91	0.92	0.92	**0.93**	**0.93**			
support	**0.90**	**0.90**	0.87	**0.90**	**0.90**	**0.90**	**0.90**			
surgery	**0.98**	0.94	0.96	**0.98**	**0.98**	0.97				
transient	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**				
transport	0.93	0.93	0.92	0.93	0.93	0.93	**0.94**			
ultrasound	0.84	0.84	0.84	0.85	0.84	0.88	0.88	**0.92**		0.88
variation	0.80	0.88	0.80	0.86	0.82	**0.94**	0.89			
weight	0.47	0.74	0.72	0.71	0.75	0.82	0.78	**0.83**	0.68	0.78
white	0.49	0.78	0.76	0.74	0.69	**0.81**	0.73	0.79	0.62	0.76

Table 4.18: Overall Results of K-CUI and Related Work

	Base-line	K-CUI Naive Bayes		K-CUI SVMs		Stevenson et. al.,	Savova et. al.,	Joshi et. al.,	Leroy and Rindflesch,	Liu et. al.,
		abst.	sent.	abst.	sent.	2008	2008	2005	2004	2004
NLM-WSD dataset	0.79	0.85	0.84	0.85	0.84	**0.88**				
Savova subset	0.76	0.83	0.81	0.83	0.81	**0.86**	**0.86**			
Joshi subset	0.68	0.77	0.75	0.77	0.74	0.80	**0.81**	0.79		
Leroy subset	0.68	0.75	0.72	0.74	0.72	**0.80**	0.79	0.77	0.66	
Liu subset	0.72	0.82	0.80	0.81	0.79	0.85	0.85	0.85		**0.86**

with this type of evaluation is that the features in the feature set are extracted from the test portion of the data as well as the training which increases the overall accuracy of the results. The characteristics of this type of evaluation is known in the WEKA community and has previously been discussed in their mailing list[1] . K-CUI does not use the 10-fold cross validation method in WEKA for this reason. The results reported by K-CUI use the internal cross validation method which only extracts the features from the training data at each fold rather than the entire dataset.

The method proposed by [Joshi et al., 2005] is the same method used above in Section 4.3 which comparing K-CUI with the general English feature unigrams. The method proposed by [Joshi et al., 2005] use the general English feature unigrams that occur in the same abstract as the target word using a frequency cutoff of four and the SVM algorithm. The evaluate their method using the 10-fold cross validation option provided by the WEKA data mining package on the Joshi subset of the NLM-WSD dataset. The results reported by [Joshi et al., 2005] are higher than those previously reported as shown in Table 4.19 due to the method of evaluation.

Table 4.19: Unigram Results using the Joshi Subset

Cutoff	0	2	4
Unigrams using 10-fold cross validation			0.77
Unigrams using WEKA 10-fold cross validation	0.76	0.74	0.66
[Joshi et al., 2005]			0.77

The overall results reported by [Joshi et al., 2005] show that their method obtains

[1] The "attribute selection and cross validation" thread of the WEKA mailing list located at https://list.scms.waikato.ac.nz/pipermail/wekalist/2008-June.txt

a higher disambiguation accuracy than K-CUI when using the WEKA 10-fold cross validation evaluation. The authors report an overall accuracy of 79% on the Joshi subset, a 77% on the Leroy subset and an 85% on the Liu subset. The K-CUI results obtain an overall accuracy of 77% on the Joshi subset, a 74% on the Leroy subset and a 81% on the Liu subset.

The method proposed by [Leroy and Rindflesch, 2004] use the semantic types of the terms in the same sentence as the target word along with the general English features: part-of-speech of the target word and whether the target word is the main word in its phrase. The authors used the Naive Bayes algorithm in their supervised method and report an overall accuracy of 66% on the Leroy subset. The K-CUI results using Naive Bayes and the feature set containing the CUIs in the same sentence as the target word obtain an overall accuracy of 72%. The difference in the results is statistically significant. The results indicate that the using the CUIs of the terms surrounding the target word are a better indicator of the concept of the target word than the semantic types of the surrounding words. Semantic types are a broad categorization of CUIs, currently, there exist 135 semantic types and approximately 1.5 million CUIs in the 2008AB version of the UMLS indicating that CUIs are a finer-grained feature than semantic types.

The method proposed by [Stevenson et al., 2008] uses MSH headings assigned to Medline abstracts, collocations, bigrams and unigrams as features into their supervised system. MSH headings are concepts from the Medical Subject Heading (MSH) vocabulary which are manually assigned to biomedical citations in PubMed for indexing purposes. The NLM-WSD dataset consists of abstracts from PubMed where each abstract contains at least one MSH heading. [Stevenson et al., 2008] report that using the feature set containing CUIs that exist in the same abstract as the target word obtain a higher disambiguation accuracy than using just the MSH headings. The combination though of MSH headings and the general English features returns a higher accuracy than just using the CUIs or the combination of CUIs and the general English features, which suggests that the MSH headings and general English features contain more complementary information than CUIs and the general English features. The reason may be that MSH headings contain less noise than CUIs. MSH headings are manually assigned by humans whereas CUIs are automatically assigned by MetaMap which does

not attempt to disambiguate terms which map to more than one CUI. The disadvantage of using MSH headings as features is that they are only available as a feature when disambiguating words in Medline citations. This is not the case with CUIs which can be used as feature to disambiguate words in any text.

The results show that using CUIs in the same abstract as the target word obtain a higher disambiguation accuracy than using the semantic types and a comparable accuracy to using unigrams. The results also show that CUIs obtain a higher disambiguation accuracy than using MSH headings but a lower accuracy when the MSH headings are combined with general English features.

4.5 Error Analysis

This section conducts an error analysis of the results obtained by the individual target words. Table 4.20 shows the accuracy of the majority sense baseline results, the Naive Bayes results when using no frequency cutoff (Accuracy), the difference between these results and the baseline (Difference), the SVM results when using no frequency cutoff (Accuracy) and the difference between these results and the baseline (Difference). The table is ranked based on the majority sense baseline which is high for most of the target words, only 15 out of the 50 target words have a majority sense less than 65% and only 24 of them have a majority sense less than 82%.

The results show that as the majority sense increases the difference between the baseline and the results decreases. There are 26 target words whose majority sense is equal to or greater than 82%, and the maximum difference between any of these results and the baseline is four percentage points with the average difference being zero for both the SVM and Naive Bayes.

The are 24 target words whose majority sense baseline is less than 82% and the average difference for these target words is 15 percentage points for both the Naive Bayes and SVM. There existed only seven target words out of these 24 that had a difference of five percentage points or less for either the SVM or Naive Bayes results:

- adjustment
- blood pressure
- determination

Table 4.20: Comparison of K-CUI Results to the Majority-sense Baseline

		Naive Bayes		SVM	
	Baseline	Accuracy	Difference	Accuracy	Difference
nutrition	0.45	0.47	0.02	0.42	-0.03
weight	0.47	0.74	0.27	0.71	0.24
white	0.49	0.78	0.29	0.74	0.25
sensitivity	0.49	0.92	0.43	0.87	0.38
evaluation	0.50	0.69	0.19	0.74	0.24
mosaic	0.52	0.78	0.26	0.76	0.24
repair	0.52	0.91	0.39	0.84	0.32
blood pressure	0.54	0.54	0.00	0.51	-0.03
man	0.58	0.83	0.25	0.84	0.26
immunosuppression	0.59	0.80	0.21	0.87	0.28
radiation	0.61	0.84	0.23	0.86	0.25
adjustment	0.62	0.65	0.03	0.65	0.03
degree	0.63	0.81	0.18	0.82	0.19
growth	0.63	0.72	0.09	0.70	0.07
scale	0.65	0.77	0.12	0.75	0.10
lead	0.71	0.90	0.19	0.93	0.22
fat	0.71	0.74	0.03	0.78	0.07
failure	0.71	0.68	-0.03	0.71	0.00
japanese	0.73	0.78	0.05	0.78	0.05
discharge	0.74	0.93	0.19	0.92	0.18
determination	0.79	0.79	0.00	0.84	0.05
variation	0.80	0.88	0.08	0.86	0.06
sex	0.80	0.88	0.08	0.88	0.08
implantation	0.81	0.94	0.13	0.92	0.11
extraction	0.82	0.84	0.02	0.84	0.02
fit	0.82	0.85	0.03	0.86	0.04
mole	0.83	0.84	0.01	0.86	0.03
ultrasound	0.84	0.84	0.00	0.85	0.01
pathology	0.85	0.85	0.00	0.83	-0.02
depression	0.85	0.83	-0.02	0.84	-0.01
cold	0.86	0.89	0.03	0.88	0.02
culture	0.89	0.91	0.02	0.89	0.00
reduction	0.89	0.89	0.00	0.89	0.00
support	0.90	0.90	0.00	0.90	0.00
condition	0.90	0.90	0.00	0.90	0.00
glucose	0.91	0.90	-0.01	0.91	0.00
strains	0.92	0.92	0.00	0.92	0.00
ganglion	0.93	0.94	0.01	0.93	0.00
transport	0.93	0.93	0.00	0.93	0.00
frequency	0.94	0.94	0.00	0.95	0.01
pressure	0.96	0.95	-0.01	0.96	0.00
resistance	0.97	0.96	-0.01	0.97	0.00
surgery	0.98	0.94	-0.04	0.98	0.00
inhibition	0.98	0.98	0.00	0.98	0.00
secretion	0.99	0.99	0.00	0.99	0.00
energy	0.99	0.99	0.00	0.99	0.00
transient	0.99	0.99	0.00	0.99	0.00
single	0.99	0.98	-0.01	0.99	0.00
fluid	1.00	0.98	-0.02	1.00	0.00
association	1.00	0.98	-0.02	1.00	0.00
Overall Accuracy	0.78	**0.85**		**0.85**	

- failure
- fit
- japanese
- nutrition

The hypothesis is that the possible concepts of these target words are too fine-grained for the system to distinguish between them. The possible concepts for each of the seven target words are:

- adjustment
  - Individual Adjustment [C0376209]
  - Adjustment Action [C0456081]
  - Psychological adjustment [C0683269]
- blood pressure
  - Blood Pressure [C0005823]
  - Blood Pressure Determination [C0005824]
  - Arterial pressure [C0428878]
- determination
  - Adjudication [C0680730]
  - Determination [C0243075]
- failure
  - Failure [C0699796]
  - Failure, NOS [C0231174]
- fit
  - Seizures [C0036572]
  - Fit and well [C0424576]
- japanese
  - Japanese language [C0376247]
  - Japanese Population [C0022342]
- nutrition
  - Nutrition [C0392209]

  - Science of nutrition [C0028707]
  - Feeding and dietary regimes [C0600072]

[Leroy and Rindflesch, 2005] show that the target word *blood pressure* is not ambiguous in normal English usage but is ambiguous in the UMLS. The first concept (C0005823) refers to the entity of blood pressure itself. The second (C0005824) refers to the act of taking blood pressure and the third (C0428878) refers to the actual pressure. The fine-granularity between the concepts make it difficult to distinguish between the concepts given that the context that they are used in will be quite similar.

The authors also state that the three concepts for the target word *adjustment* are not distinguishable. The first (C0376209) and third (C0683269) concepts refer to the psychological state of an individual and the semantic type for each of them is "Individual Behavior". The definitions of each of concepts are also very similar:

- Individual Adjustment [C0376209]
  - Conceptually broad term referring to a state of harmony between internal needs and external demands and the processes used in achieving this condition. Use a more specific term if possible. Differentiate from ADAPTATION, which refers to physiological or biological adaptation.
- Psychological adjustment [C0683269]
  - A state of harmony between internal needs and external demands and the processes used in achieving this condition.

This analysis can also be done for the target word *nutrition*. There are three possible concepts for *nutrition*, Nutrition [C0392209], Science of Nutrition [C0028707] and Feeding and Dietary Regimes [C0600072], with two having a corresponding definition in the UMLS:

- Nutrition [C0392209]
  - State of the body in relation to the consumption and utilization of nutrients.
- Science of nutrition [C0028707]
  - The study of NUTRITION PROCESSES as well as the components of food, their actions, interaction, and balance in relation to health and disease.

The first concept (C0392209) involves the state of an individual with relation to the consumption of food and utilization of the nutrients in the food by the body. The second concept (C0028707) is the study of how the nutrients are utilized and the third concept (C0600072) is the act of consumption. The definitions indicate that the distinction between these concepts is very fine-grained.

The analysis of these three target words is conducted using human intuition on the fine-granularity between the concepts based. A second analysis is conducted by comparing the context in which the concepts are used to determine if the context surrounding the possible concepts for these target words is similar.

In this analysis, the number of times a feature-concept pair occurs is occurs in an instance and the overlap is calculated by tallying up the number of times it occurs overall and the number of times it occurs with more than one of the possible concepts. The CUIs in instances assigned "None" are included because they are treated like an additional concept by K-CUI.

Table 4.21 shows the target word, the overlap, the overlap percentage and the accuracy obtained for that target word using the SVM algorithm ranked based on the percentages. For example, the target word *pressure* has 2,664 distinct CUIs in the feature set with 118 of those CUIs existing in the same abstract as at least two of its possible concepts, the percentage of overlap is 0.04, and it has an accuracy of 96%.

The results show that the target word's accuracy is correlated with the percentage of overlap. For example, target word $fluid$ does not have any overlapping CUIs and obtains an accuracy of 100% while the target word *nutrition* has 74% of the CUIs in the feature set exist in the same abstract as at least two its possible concepts and obtains an accuracy of 42%.

The rankings of the seven target words from the previous analysis are:

- (23) determination
- (28) fit
- (32) japanese
- (39) failure
- (48) adjustment
- (49) blood pressure
- (50) nutrition

Table 4.21: Overlap of CUIs in the NLM-WSD Dataset

Rank	Target Word	Overlap	Percentage	Accuracy
1	fluid	0/2867	0.00	1.00
2	association	0/2833	0.00	1.00
3	transient	46/2848	0.02	0.99
4	single	68/2824	0.02	0.99
5	energy	104/2561	0.04	0.99
6	inhibition	136/2645	0.05	0.98
7	secretion	148/2536	0.06	0.99
8	surgery	190/2535	0.07	0.98
9	resistance	208/2623	0.08	0.97
10	pressure	236/2664	0.09	0.96
11	ganglion	250/2410	0.10	0.93
12	transport	481/2601	0.18	0.93
13	culture	540/2813	0.19	0.89
14	frequency	578/2771	0.21	0.95
15	strains	542/2254	0.24	0.92
16	glucose	636/2566	0.25	0.91
17	condition	678/2699	0.25	0.90
18	reduction	783/3052	0.26	0.89
19	support	740/2885	0.26	0.90
20	depression	628/2128	0.30	0.84
21	variation	852/2776	0.31	0.86
22	cold	906/2767	0.33	0.88
23	**determination**	912/2774	0.33	0.84
24	pathology	886/2645	0.33	0.83
25	implantation	816/2485	0.33	0.92
26	extraction	912/2646	0.34	0.84
27	ultrasound	858/2328	0.37	0.85
28	**fit**	988/2647	0.37	0.86
29	sex	968/2584	0.37	0.88
30	discharge	999/2663	0.38	0.92
31	mole	973/2575	0.38	0.86
32	**japanese**	941/2365	0.40	0.78
33	lead	1278/2861	0.45	0.93
34	sensitivity	1259/2737	0.46	0.87
35	evaluation	1274/2702	0.47	0.74
36	man	1228/2610	0.47	0.84
37	degree	1420/2959	0.48	0.82
38	scale	1254/2545	0.49	0.75
39	**failure**	1435/2893	0.50	0.71
40	growth	1414/2678	0.53	0.70
41	immunosuppression	1276/2412	0.53	0.87
42	radiation	1376/2525	0.54	0.86
43	white	1509/2642	0.57	0.74
44	fat	1362/2407	0.57	0.78
45	repair	1479/2559	0.58	0.84
46	mosaic	1300/2184	0.60	0.76
47	weight	1724/2720	0.63	0.71
48	**adjustment**	1582/2432	0.65	0.65
49	**blood pressure**	1497/2254	0.66	0.51
50	**nutrition**	2076/2451	0.85	0.42

The rankings show the target words *adjustment*, *blood pressure* and *nutrition* have the largest overlap indicating that the possible concepts of these target words is very fine-grained. The rankings of the remaining target words show that there exists a large amount of overlap between the concepts contexts indicating that there is not a clear distinction between them.

The overall conclusion of the case analysis is that the fine-granularity of some of the concepts in the NLM-WSD dataset make them difficult to disambiguate because the context surrounding the target words are similar for each of the concepts. K-CUI obtains overall disambiguation accuracy of 85% on this dataset indicating that it has the potential to perform much higher on a dataset whose concepts are not as fine-grained.

4.6 Conclusions

In this chapter, two sets of experiments were conducted to determine what CUIs to include in the feature set. The first are the windowing experiment, and the second are the cutoff experiments.

The overall results of the windowing experiment show that extracting the CUIs from the abstract containing the target word obtains a higher disambiguation accuracy than extracting the CUIs from the phrase containing the abstract. The difference in accuracy though is only four percentage points indicating that the majority of instances can be disambiguated using the two or three CUIs closest to the target word.

The overall results of the cutoff experiments show that not using cutoff obtained the highest disambiguation accuracy. The results using a frequency cutoff show that removing lower frequency CUIs reduces the overall disambiguation accuracy indicating that lower frequency features affect the supervised learning model.

The results using the MMI cutoff show that it is able to significantly reduce the amount of noise in the feature set. When using an MMI cutoff of ten, the feature set contains almost 60% fewer features going from on average 3752.646 features to 1489.92, and almost 70% less features seen in the instance going from on average 63.49 features to 18.58. The results show that the overall accuracy only decreases by one percentage point indicating that the MMI cutoff can be used to remove CUIs that are not needed in the disambiguation process while still maintaining a comparable accuracy.

The results using the semantic similarity cutoff also show that it is able to significantly reduce the amount of noise in the feature set. The feature set when using the semantic similarity measure proposed by [Wu and Palmer, 1994] with a cutoff 0.1 contains almost 70% less features going from 3752.64 to 1026.38, and almost 60% less features going from an average 63.49 features to 25.62. The overall accuracy is only two percentage points lower than not using a cutoff indicating that the semantic similarity cutoff can be used to remove CUIs that are not needed in the disambiguation process while still maintaining a comparable accuracy.

A comparative analysis was conducted between the general English feature set containing unigrams and K-CUI. The overall results show that K-CUI and the unigram feature obtain the same overall disambiguation accuracy when not using a frequency cutoff but when using a frequency cutoff the accuracy of the unigram results decreases at a greater rate than the CUI results indicating that CUIs are a more robust feature for disambiguation.

A comparative analysis was also conducted between K-CUI and previously proposed supervised WSD methods that have been evaluated on the NLM-WSD dataset. The overall results show that K-CUI either obtains a higher disambiguation accuracy or is comparable to results of the previous except for the method proposed by [Stevenson et al., 2008] which uses a combination of MSH headings and general English features. The disadvantage of MSH headings though is that they are only available as a feature when disambiguating words in Medline abstracts whereas the CUI feature used by K-CUI can be used to disambiguate words in any biomedical text.

Lastly, an analysis of the NLM-WSD dataset was conducted showing that many of the possible concepts of the target words are fine-grained making it difficult to disambiguate between them. Regardless, K-CUI obtains overall disambiguation accuracy of 85% on this dataset indicating that it has the potential to perform much higher on a dataset whose concepts are more clearly distinct.

The overall conclusion of these experiments is that using the CUIs of the words surrounding the target word are a good indicator of the target word's concept. K-CUI can be used to disambiguate words in other biomedical text not just Medline abstracts and the results indicate that on other datasets the overall disambiguation accuracy will increase. A more detailed analysis of these results is in Chapter 9.

Chapter 5

A-CUI

This chapter describes a proposed knowledge-based WSD method, called A-CUI, that uses information from the UMLS and MetaMap mapped text to disambiguate words in biomedical text. In this method, a first or second-order test vector is created using the context surrounding the target word, and corresponding first or second-order concept vectors are created for each possible concept of the target word using information about the concept from the UMLS or MetaMap mapped text. A measure is used to quantify the distance between the test vector and each of the possible concept vectors. The concept whose vector is closest to the test vector is assigned to the target word. The novelty of A-CUI is the creation and development of a knowledge-based vector method which determines the correct concept of a target using information from the UMLS and MetaMap mapped text.

The following sections discuss the i) motivation behind A-CUI, ii) the algorithm used to implement A-CUI and iii) the actual A-CUI implementation.

5.1 Motivation

The concept vector for A-CUI is created using information from an outside knowledge source. This method is different from other methods that create a context vector by taking the centroid of a set of training vectors that have been manually annotated or automatically labeled with that concept. A-CUI does not rely on examples from a training data but descriptive information about a concept extracted from the UMLS or

MetaMap mapped text to provide a contextual representation of the concept.

In this method, A-CUI extracts contextual information about a concept's CUI and then the concept vectors are created using the words from the contextual information as described in Section 2.1. This dissertation investigates three different types of contextual representations: CUI definitions, terms associated with the CUI, and frequently occurring words surrounding a CUI or its associated terms in MetaMap mapped text.

The assumption behind this method is that the information extracted from the UMLS or MetaMap mapped text provides enough contextual information about how a concept is used in biomedical text to distinguish between them when presented with an instance containing the target word.

The following section in this chapter discusses the algorithm describing the proposed knowledge-based method. The next section describes the actual implementation of this method and the options available to creating the contextual representation of a target word's possible concepts.

5.2 Algorithm

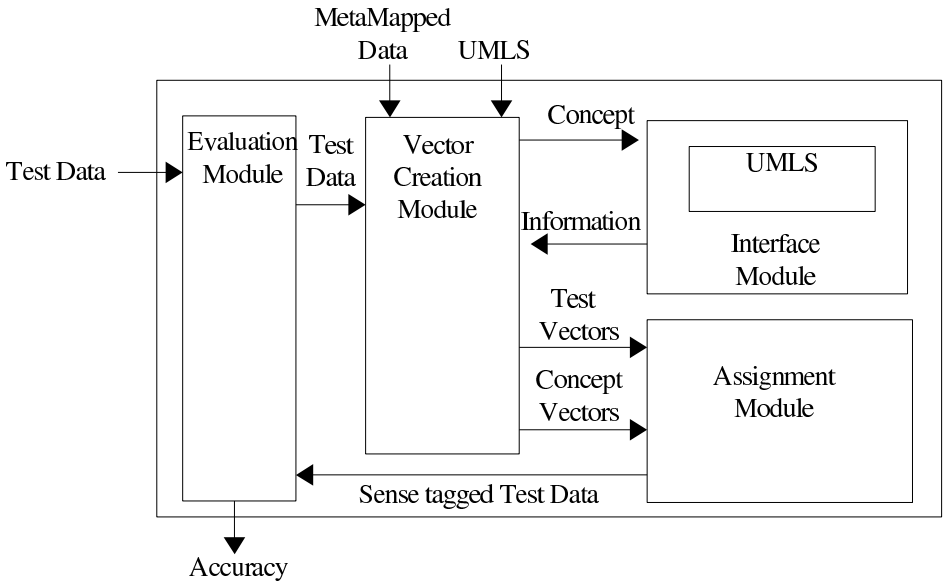

Figure 5.1: A-CUI Algorithm

This section describes the A-CUI algorithm shown in Figure 5.1. A-CUI is composed of four modules: the evaluation module, the vector creation module, the interface module and the assignment module. The evaluation module sets up the testing framework in order to determine the accuracy of the A-CUI experiments which is necessary for evaluation purposes. The vector creation module creates the feature set, extracts the contextual representation of a concept from the interface module and creates the test and concept vectors, the assignment module assigns concepts to the instances in the test data. The pseudocode for the main driver program of A-CUI is in Algorithm 5.1.

The main driver program takes five variables as input: Annotated Data, Unannotated Data, Vector Type, Representation Type and the Metric. The Annotated Data is the data in which A-CUI is evaluated on. It contains a set of instances containing a target word whose concept has been manually identified. The annotations are strictly for evaluation purposes only. The Unannotated data is the data which is later processed by MetaMap to extract information about a concept's CUI to create its contextual representation. The Vector Type contains the information required to specify whether the test and concept vectors are to be either first-order or second-order vectors. The Representation Type contains the information required to specify which information should be used as the contextual representations of the possible concepts. The Metric specifies which metric to calculate the distance between the test and concept vectors.

As the pseudocode of the main A-CUI drivers shows, there are four main steps. In step 1, the *Evaluation Module* creates the test data; the pseudocode for this module is shown in Algorithm 5.2. In this module, the RemoveAnnotations() function removes the concepts from the manually annotated training data which are there only for evaluation purposes.

In step 2, the *Vector Creation Module* creates the test and concept vectors; the pseudocode for this module is shown in Algorithm 5.3. The CreateVectors() function takes the test data, training data, vector type and representation type as input and returns either first-order or second-order test and concept vectors depending on the vector type. In this function, the feature set is created by extracting all of the words in the training data, the target word is identified in the test data, and all of the possible concepts of the target word are obtained. A test vector is then created using the information in the instance, and a concept vector is created for each possible concept

Algorithm 5.1 A-CUI Pseudocode

procedure A-CUI(*AnnotatedData*, *UnannotatedData*, *VectorType*, *RepresentationType*, *Metric*)
comment: Step 1:Remove Annotations from Manually Annotated Data
TestData = RemoveAnnotations(*AnnotatedData*)
comment: Step 2: Create Concept and Test Vectors
(*ConceptVectors*, *TestVectors*) =
    CreateVectors(*TestData*, *UnannotatedData*, *VectorType*, *RepresentationType*)
comment: Step 3: Assign Concepts to Test Vectors
ConceptTaggedTestVectors = AssignConcepts(*TestVectors*, *ConceptVectors*, *Metric*)
comment: Step 4: Calculate Accuracy of the System
Accuracy = CalculateAccuracy(*ConceptTagged TestVectors*, *AnnotatedData*)
print Accuracy

Algorithm 5.2 Evaluation Module Pseudocode

function RemoveAnnotations(*Data*)
UnnannotatedData = RemoveConcepts(*Data*)
return (*UnannotatedData*)

function CalculateAccuracy(*ConceptTagged TestVectors*, *AnnotatedData*)

Correct = GetNumberCorrect(*ConceptTaggedTestVectors*, *AnnotatedData*)
Wrong = GetNumberWrong(*ConceptTaggedTestVectors*, *AnnotatedData*)
Accuracy = *Correct* / (*Correct* + *Wrong*)
return (*Accuracy*)

Algorithm 5.3 Vector Creation Module Pseudocode

```
function CreateVectors(TestData, TrainingData, VectorType, RepresentationType)
 FeatureSet = ExtractWords(TrainingData)
 TargetWord = GetTargetWord(TestData)
 PossibleConcepts = GetConcepts(TargetWord)
 for each Concept ∈ PossibleConcepts
 { Representation = GetRepresentation(Concept, TrainingData, RepresentationType)
 { if VectorType = First − Order
 {   then Vector = CreateFirstOrderVector(FeatureSet, Representation)
 {   else Vector = CreateSecondOrderVector(FeatureSet, Representation, TrainingData)
 { comment: Add the Vector to an array of Concept Vectors to be returned
 { ConceptVectors ← Vector
 for each Instance ∈ TestData
 { if VectorType = First − Order
 {   then Vector = CreateFirstOrderVector(FeatureSet, Instance)
 {   else Vector = CreateSecondOrderVector(FeatureSet, Instance, TrainingData)
 { comment: Add the Vector to an array of Test Vectors to be returned
 { TestVectors ← Vector
 return (ConceptVectors, TestVectors)

function CreateFirstOrderVector(FeatureSet, Instance)
 comment: Create vector where each element is a feature in the Feature Set
 Vector = InitializeVector(Vector)
 for each Feature ∈ Vector
 { if Feature ∈ Instance
 {   then Vector[Feature] = 1
 {   else Vector[Feature] = 0
 return (Vector)

function CreateSecondOrderVector(FeatureSet, Instance, Data)
 comment: Create First order vector for each word in the instance and added to the WordVectors array
 for each Word ∈ Instance
 { WordVector = CreateFirstOrderVector(FeatureSet, Data)
 { WordVectors ← WordVector
 comment: Average WordVectors creating Second order vector
 SecondOrderVector = AverageVectors(WordVectors)
 return (SecondOrderVector)
```

using the contextual representation obtained from the GetContext() function in the *Interface Module* shown in Algorithm 5.4.

Algorithm 5.4 Interface Module Pseudocode

function GETREPRESENTATION(*Concept*, *MetaMappedData*, *RepresentationType*)
 $Type = RepresentationType[0]$;
 if $Type = Definition$
  then $\begin{cases} Relations = RepresentationType[1] \\ Representation = \text{GETDEFINITIONS}(Concept, Relations) \end{cases}$
  else if $Type = Terms$
  then $\begin{cases} TermType = RepresentationType[1] \\ Representation = \text{GETTERMS}(Concept, TermType) \end{cases}$
  else if $Type = MetaMap$
  then $\begin{cases} Feature = RepresentationType[1] \\ X = RepresentationType[2] \\ MetaMappedData = \text{METAMAP}(TrainingData) \\ Representation = \text{GETMETAMAPPEDCONTEXT}(Concept, Feature, X, MetaMappedData) \end{cases}$
 return (*Representation*)

The *Interface Module* directly extracts the information from the UMLS; the pseudocode for this module is shown in Algorithm 5.4 and 5.5. The GetDefinitions() function extracts the definition of the concept or the definition of its related concepts from the UMLS. The GetTerms() function extracts a concepts preferred or associated terms from the UMLS. The GetMetaMapContext() function extracts the top X most frequent words that surround a concepts CUI in MetaMap mapped text, or the top X most frequent words that surround a concepts associated terms.

In step 3, the *Assignment Module* assigns a concept to each of the test vectors; the pseudocode for this is shown in Algorithm 5.6. The AssignConcepts() function takes the test and concept vectors as input, a metric such as the Cosine Measure is calculated between a test vector and each of the concept vectors, and the concept whose vector is closest to the test vector is assigned to the target word. This is done for each of the test vectors creating a set of concept tagged test vectors.

Algorithm 5.5 Interface Module Pseudocode (Continued)

function GETDEFINITIONS($Concept$, $Relations$)
 INITIALIZEDEFINITIONS($Definitions$)
 for each $Relation \in Relations$
 if $Relation = CUI$
 then CUI = GETCUI($Concept$)
 $Definitions = Definitions +$ GETCUIDEFINITION(CUI)
 else CUI = GETCUI($Concept$)$Rels$ = GETRELATIONS($Relation$, CUI)
 for each $Rel \in Rels$
 $Definitions = Definitions +$ GETCUIDEFINITION(Rel)
 return ($Definitions$)

function GETTERMS($Concept$, $TermType$)
 CUI = GETCUI($Concept$)
 if $TermType = AssociatedTerms$
 then $Terms$ = GETASSOCIATEDTERMS(CUI)
 else $Terms$ = GETPREFERREDTERM(CUI)
 return ($Terms$)

function GETMETAMAPPEDCONTEXT($Concept$, $Feature$, X, $MetaMappedData$)
 comment: Extract the top X most frequent words surrounding the CUI or its associated terms
 CUI = GETCUI($Concept$)
 if $Feature = CUIs$
 then $Words$ = EXTRACTWORDS($MetaMappedData$, CUI, X)
 else $AssociatedTerms$ = GETASSOCIATEDTERMS(CUI)
 $Words$ = EXTRACTWORDS($MetaMappedData$, $AssociatedTerms$, X)
 return ($Words$)

Algorithm 5.6 Assignment Module Pseudocode

```
function AssignConcepts(TestVectors, ConceptVectors, Metric)
 for each TestVector ∈ TestVectors
 { MinimumDistance = ∞
 { for each ConceptVector ∈ ConceptVectors
 { { comment: The distance is obtained between the test and concept vector using a specified metric
 { Distance = GetDistance(TestVector, ConceptVector, Metric)
 { if MinimumDistance > Distance
 {   then { MinimumDistance = Distance
 {        { AssignedTestVector = AssignConceptToVector(ConceptVector, TestVector)
 { comment: Add the Assigned Test Vector to an array of Assigned Test Vectors to be returned
 { AssignedTestVectors ← AssignedTestVector
 return (AssignedTestVectors)
```

In step 4, the *Evaluation Module* then calculates the accuracy of the method; the pseudocode for this module is shown in Algorithm 5.2. The CalculateAccuracy() function takes the concept tagged test vectors and the manually assigned data as input and calculates the accuracy of the assignments. The following section describes the actual implementation details of this algorithm and various metric and contextual representation options available.

5.3 System

This section discusses the implementation details of A-CUI which is shown in Figure 5.2. A-CUI takes test data containing instances of a target word assigned their appropriate concept as input. The concept information is included in the test data for evaluation purposes only; manually annotated training is not required for this method.

The evaluation module removes the concept information from the test data and sends it to the vector creation module. The vector creation program takes the now-untagged test data as input and creates a first-order or second-order test vector for each instance

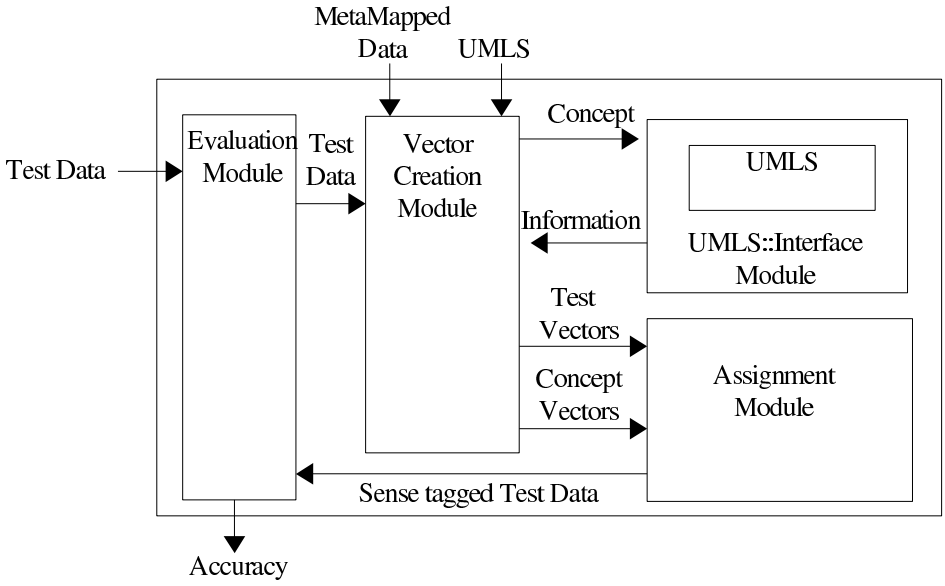

Figure 5.2: A-CUI System

in the test data, and corresponding first or second-order concept vector for each possible concept of the target word. The program used to create these vectors originates from the clustering word sense discrimination package, SenseClusters[1] .

The feature set used to create the vectors contains words that occur more than five times in the 2005 Medline baseline and do not occur in a stoplist[2]. The frequency cutoff and stoplist are used to reduce the amount of noise that can exist in the feature set.

The test vector is created using the words surrounding the target word, which is a common approach in WSD methods and is used in the clustering WSD method and the k-nearest neighbor algorithm discussed in Section 2.2.3. The concept vectors are created using the following information extracted from the UMLS and the 2005 Medline baseline:

- its UMLS CUI definition
- the UMLS CUI definitions of its related concepts
- the UMLS CUI's preferred term
- the UMLS CUI's associated terms

[1] For a more detailed description SenseClusters see Section 2.4.3
[2] Appendix F contains the stoplist used by A-CUI

- frequently occurring words surrounding the concept's CUI in the MetaMapped 2005 Medline baseline
- frequently occurring words surrounding a CUI's associated terms in the MetaMapped 2005 Medline baseline

This information is extracted using the freely available, platform independent, open source module created by [McInnes et al., 2009] called UMLS::Interface[3].

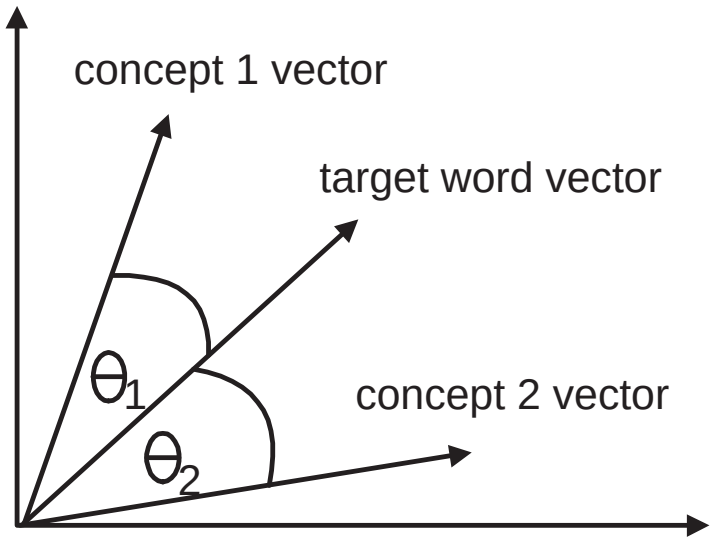

Figure 5.3: A-CUI Algorithm

The A-CUI algorithm module takes the test and concept vectors from the vector creation program as input, and, for each of the test vectors, calculates the distance between the test vector and each possible concept vectors, as seen in Figure 5.3. The concept whose vector is closest to the test vector is assigned to the target word. The A-CUI algorithm calculates this distance using one of three vector metrics:

- Cosine Measure
- Dice Coefficient
- Euclidean Distance

[3] UMLS::Interface can be downloaded at http://search.cpan.org/dist/UMLS-Interface/

The Euclidean distance is the sum of the distance between two points in two vectors, X and Y. The closer two vectors are the lower their distance score. It is mathematically defined as:

$$distance = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}. \tag{5.1}$$

where n is the size of the vectors.

The Cosine measure is a measure of similarity between two vectors, X and Y, of n dimensions by finding the cosine of the angle between them. The cosine ranges from -1 (exactly opposite) to 1 (exactly the same). It is mathematically defined as:

$$cosine = \frac{X \cdot Y}{\|X\|\|Y\|}. \tag{5.2}$$

Another vector similarity measure is the Dice Coefficient. It determines the similarity between two vectors, X and Y, by counting the number of times both vectors contain the same element and dividing it by the sum of the number of elements in the vectors. The Dice Coefficient ranges from 0 to 1 with 1 indicating that the vectors are the same. It is mathematically defined as:

$$dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5.3}$$

The remainder of this section discusses the available options A-CUI has to represent the context of a concept in order to create its associated concept vector.

5.3.1 UMLS CUI Definitions

This section describes using the definition of a concept's CUI to represent its context. Concept definitions have previously been used to represent the context of a concept for the task of calculating the semantic relatedness between two concepts but not for the task of WSD.

The semantic relatedness measure, WordNet::Similarity::vector[4], proposed by [Patwardhan, 2003] calculates the semantic relatedness between two concepts in WordNet by creating a second-order concept vector for each concept using its definition (or gloss as it is called in WordNet) and the definition of its related concepts as context.

[4] http://sourceforge.net/projects/wn-similarity/

The cosine measure is then used to calculate the angle between the two vectors which is used to quantify their similarity. The assumption behind the Cosine Measure is that the definitions provide contextual information about a concept and similar concepts will have similar contextual information.

A-CUI uses the UMLS CUI definitions to provide a context for each of the possible concepts of a target word. Consider the term *culture* which has two possible concepts in the NLM-WSD dataset, each of those concepts has at least one corresponding definition in the UMLS although there do exist some CUIs that do not.

- Anthropological Culture [C0010453]
  - A collective expression for all behavior patterns acquired and socially transmitted through symbols. Culture includes customs, traditions, and language.
  - A pattern of learned beliefs, values, and behavior that are shared within a group. It includes language, styles of communication, practices, customs, and views on roles and relationships.
- Laboratory Culture [C0430400]
  - Any laboratory procedure for growing microorganisms.

The features for the concept vectors contain the words in the 2005 Medline baseline and the elements for a first-order concept vector are either a one or zero indicating whether or not a feature exists in the concepts definition while the second-order context vector is the average of each of the first-order vectors of the content words in the definition. The first-order vectors provide a representation of the definition while the second-order vectors provide a representation of the context in which the words in the definition are used.

A-CUI also allows for combinations of the following related concepts definitions to be included as context:

- PAR/CHD: parent/child
- RB/RN: broader/narrower than
- SY: source asserted synonymy
- RO: has a relationship other than synonymous, narrower, or broader
- RL: concepts are similar or "alike".

- RQ: related and possibly synonymous
- SIB: sibling
- AQ: allowed qualifier
- QB: can be qualified by
- RQ: related and possibly synonymous
- RU: related but unspecified
- XR: not related

If the definition of a CUI's relation is used, the elements in the first-order concept vector are either a one or a zero indicating whether or not a word in the feature set occurs in the definition of its related CUI, and the second-order concept vector is the average of the first-order vectors of the content words in the definition of the concept's related CUI.

This dissertation investigates the PAR, CHD, SY and SIB relations. The remainder of the section discusses these four relations. The PAR/CHD relation in the UMLS contains hierarchical *is-a* and *part-of* relations that are explicitly expressed by a source. An *is-a* relation is where one class is a subclass of another class, and a *part-of* relation is where one class is part of another class.

Consider again the target word *culture* which has two possible concepts: Anthropological Culture [C0010453] and Laboratory Culture [C0430400]. Anthropological Culture [C0010453] has the parent Sociology/Anthropology [C0178848] which has the definition:

- Header term for two closely related sciences; sociology is the social science dealing with group relationships, patterns of collective behavior, and social organization; anthropology is the science devoted to the comparative study of man.

While Laboratory Culture [C0430400] has the parent Microbiological Technics [C0025951] which has the definition:

- Techniques used in microbiology.

The assumption is that the definition of concept's parents provides broad general contextual information about a concept.

The children of a concept provide the opposite type of information because each of the children are special cases of their parents. For example, one of the children of Anthropological Culture [C0010453] is Ceremonial Behaviors [C0007825] which has the definition:

- A series of actions, sometimes symbolic actions which may be associated with a behavior pattern, and are often indispensable to its performance.

While one of the children of Laboratory Culture [C0430400] is Culture Setup [C1511554] which has the definition:

- The preparation work required for doing cell or tissue culture. This includes preparation of the medium and of the growth support material or growth chamber.

The assumption is that the aggregation of the definitions of a concepts children will provide enough contextual information to disambiguate between the two concepts.

The SIB relation links concepts that share the same parent in the UMLS. Consider the target word *ganglion* which has two possible concepts: Benign Cystic Mucinous Tumour [C0085648] and Ganglia [C0017067]. One of the siblings of Benign Cystic Mucinous Tumour [C0085648] is Lymph Cyst (C0024248) which has the definition:

- Cystic mass containing lymph from diseased lymphatic channels or following surgical trauma or other injury.

While one of the siblings of Ganglia [C0017067] is Sense Organs [C0036665] which has the following definition:

- Structure which is a receptor for external or internal stimulation.

The assumption is that siblings will contain similar contextual information.

The SY relation links concepts in which some source in the UMLS states that they are synonyms but they are considered separate concepts by the UMLS editors. There are very few SY relations in the UMLS. The assumption is that this may provide additional information to at least one of the possible concepts of a target word to help make a distinction between them. Consider the target word *man* which has three possible concepts: Male [C0024554], Men [C0025266] and Homo Sapiens [C0086418]. The concept Homo Sapiens [C0086418] is linked to the concept Women [C0043210] which has the definitions:

- Human adult females as cultural, psychological, sociological, political, and economic entities.
- An adult, female human.

This addition information may help distinguish between concepts Male [C0024554] and Men [C0025266].

The context consists of any combination of the above definitions. This dissertation investigates using the CUI definition in conjunction with the definitions of the CUI's PAR, CHD, SIB and SY relations to create a first and second-order concept vectors.

5.3.2 UMLS CUI Terms

This section describes using the preferred or associated terms to represent the context of a concept. Term information has also previously been used in the measure to calculate the semantic relatedness between two concepts.

The semantic relatedness measure, SnoMed::Similarity::vector, proposed by [Pedersen et al., 2007] calculates the semantic relatedness between two concepts in SNOMED-CT by creating a second-order concept vector for each concept using the terms associated with a concept for in the Mayo Clinic thesaurus. The Cosine Measure is then used to calculate the angle between the two vectors and the angle is used to quantify their similarity.

Table 5.1: Associated Terms of the Target Word *Culture*

Anthropological Culture [C0010453]	Laboratory Culture [C0430400]
culturally	culture procedure
cultures	sample culture
cultural	culture
culture	sample culture procedure

A-CUI uses the preferred term of a CUI as well as its associated terms to represent the context of a possible concept. For example, consider the term *culture* which has two possible concept in the UMLS: Anthropological Culture [C0010453] and Laboratory Culture [C0430400]. "Anthropological Culture" is the preferred term for C0010453 and "Laboratory Culture" is the preferred term for C0430400. The assumption is that the

preferred terms are distinctive enough to be able to disambiguate between the possible concepts.

Associated terms are terms listed in the UMLS that have been used to describe the concept. For example, Table 5.1 shows the associated terms for the two concepts of *culture*. Although there is some overlap between the words, the assumption is that the addition of the associated terms would strengthen that distinction between the concepts.

The features for the concept vectors contain the words in the 2005 Medline baseline and the elements for a first-order concept vector are either a one or zero indicating whether or not a feature is a word in the concepts preferred or associated terms while the second-order context vector is the average of each of the first-order vectors of the words in the concepts preferred or associated terms.

5.3.3 MetaMap Mapped Text

This section describes using the frequently occurring words in the MetaMap mapped text surrounding the CUI or associated terms of a concept to represent its context. The training data consists of the 2005 Medline baseline where each term in the baseline is mapped to a CUI in the UMLS by MetaMap. The words in the same abstract as the CUI or associated terms are extracted. This dissertation investigates using the top 50 and top 100 most frequent terms.

Given that MetaMap does not perform WSD, this may lead one to think that all of the terms would be the same since MetaMap can not distinguish between the concepts. In some cases this can happen, for example, Table 5.2 shows the top ten most frequent words in the 2005 Medline baseline in the same abstract as the possible concepts of the target word *culture*; eight out of the ten words are the same. MetaMap does not map individual words to concepts in the UMLS, it maps terms. So this would be correct if the term always being mapped is *culture* but if the term is *laboratory culture* MetaMap would map it to its correct concept Laboratory Culture [C0430400]. More details on why this is the case is in Section 2.4.1.

A-CUI also has the option to use frequently occurring words that surround the associated terms of a possible concept. For example, consider the associated terms for each of the possible concepts of the target word *culture* shown in Table 5.1 in the previous section. The words in the same abstract as the associated terms are

Table 5.2: Top 10 Most Frequent Words Surrounding CUIs

Anthropological Culture [C0010453]	Laboratory Culture [C0430400]
culture	culture
cultures	cultured
cells	cells
cultured	cell
cell	human
human	growth
growth	medium
rat	rat
primary	vitro
medium	days

extracted with their frequency counts and the context used to describe the CUI consists of the words whose frequency count is above a specified threshold. This dissertation investigates using the top 50 and top 100 most frequent terms. The assumption is that the words surrounding the associated terms are able to describe the concept enough for the computer to be able to distinguish between them.

Chapter 6

A-CUI Results

This section evaluates A-CUI and discusses the results. There are four main experiments conducted. Section 6.1 investigates the metric and feature vector options available in A-CUI. A-CUI contains three metrics available to calculate the distance between the test vector and each of the possible concept vectors to determine which concept vector is the closest: Euclidean Distance, Cosine measure and Dice Coefficient. A-CUI creates the test and concept vectors as either first-order or second-order vectors. The purpose of this experiment is to determine which combination obtains the highest disambiguation accuracy.

Section 6.2 investigates using the following five contexts that use the CUI definition to create the concept vector:

- the definition of a concepts CUI
- the definition of the CUI + the parent definitions
- the definition of the CUI + the children definitions
- the definition of the CUI + the sibling definitions
- the definition of the CUI + the synonym definitions

The purpose of the definition experiments is to determine if using a CUI's definition, as well as the definition of its related concepts, provide enough contextual information to disambiguate between the possible concepts of a target word.

Section 6.3 investigates using the following contexts that use the preferred and associated terms of a CUI to create the concept vector:

- a CUIs preferred terms
- a CUIs associated terms

The purpose of the term experiments is to determine if using the terms to describe a concept provides enough distinct information to distinguish between the possible concepts of a target word.

Section 6.4 investigates using the following four contexts that use highly frequent words in the MetaMapped 2005 Medline baseline:

- 50 most frequent words in the same abstract as the possible concepts CUI
- 100 most frequent words in the same abstract as the possible concepts CUI
- 50 most frequent words in the same abstract as the terms associated with the possible concepts CUI
- 100 most frequent words in the same abstract as the terms associated with the possible concepts CUI

The purpose of the MetaMap mapped text experiments is to determine if using highly frequent words that exist in the same abstract as the CUI assigned by MetaMap or its associated terms provide enough unique contextual information in order to distinguish between the possible concepts of the target word.

Section 6.5 compares the K-CUI results to the results reported by researchers that have evaluated their knowledge-based WSD methods using the NLM-WSD dataset.

There are a few commonalities between all of the A-CUI experiments. The experiments conducted in this chapter use a subset of the NLM-WSD dataset. The instances in the dataset were manually disambiguated by annotators who assigned the target word to a concept in the UMLS (CUI) or assigned the concept as "None" if none of the possible concepts described the target word. A-CUI does not assign the concept "None" to a target word because it is unable to create a vector containing the context of None. Due to this, the subset does not contain the instances assigned "None". Table 6.1 shows the number of instances for each target word in the dataset after instances of None were removed. The target word *association* does not contain any instances assigned a UMLS concept and the target word *resistance* only contains three.

The experiments compare each of the results to two baselines: the majority sense baseline and the random baseline. The majority sense baseline is commonly used to

Table 6.1: NLM-WSD subset

target word	# Instances	target word	# Instances	target word	# Instances
adjustment	93	association	0	blood pressure	100
cold	95	condition	92	culture	100
degree	65	depression	85	determination	79
discharge	75	energy	100	evaluation	100
extraction	88	failure	29	fat	73
fit	18	fluid	100	frequency	94
ganglion	100	glucose	100	growth	100
immunosuppression	100	implantation	98	inhibition	99
japanese	79	lead	29	man	92
mole	84	mosaic	97	nutrition	89
pathology	99	pressure	96	radiation	98
reduction	11	repair	68	resistance	3
scale	65	secretion	100	sensitivity	51
sex	100	single	100	strains	93
support	10	surgery	100	transient	100
transport	94	ultrasound	100	variation	100
weight	53	white	90		

analyze the results of supervised WSD methods. The majority sense baseline is the accuracy that would be achieved by assigning every instance of the target word with the most frequent concept as assigned by the human evaluators. This baseline is what methods that do not use manually annotated training data hope to achieve but it does not always happen. The random baseline is the accuracy that would be achieved if every instance is assigned a random concept. The baseline is considered a lower bound for methods that do not use manually annotated training data.

These experiments use the pairwise t-test to determine the significance between the results. The pairwise t-test compares the accuracy of a target word from the results of one experiment with the accuracy of the same target word from the results of another experiment. The t-test tests if the sum of the change between the accuracy of the two experiments differs statistically significantly from zero, which is the same significance tests used in the K-CUI experiments.

6.1 Distance Metric and Feature Vector Results

This section discusses the metric and feature vector experiments. The purpose of these experiments is to determine which combination of options obtains the highest overall disambiguation accuracy. The results shown are the overall results of the NLM-WSD dataset; Appendix G contains a complete list of the individual target word results.

The three metrics used to determine the distance between the test and concept vector are the Euclidean Distance, Cosine measure and Dice Coefficient. The Euclidean Distance calculates the actual distance between each of the elements in the vector. The Cosine measure is a similarity measure that calculates the similarity between two vectors by calculating the angle between them. The Dice Coefficient, which is also a similarity measure, determines the similarity between two vectors by calculating the number of times the two vectors contain the same element. The purpose of this experiment is to determine which metric obtains the highest disambiguation accuracy.

In the k-Nearest Neighbor algorithm, the distance between two vectors is calculated using the Euclidean Distance. The assumption is that vectors that have a similar context will be closer together than vectors that do not. In the clustering WSD methods and the supervised method proposed by [Agirre and Martinez, 2004] the similarity is calculated using the Cosine measure. The assumption is that the angle between vectors that have a similar content will be smaller than the angle between vectors that do not. The underlying hypothesis of the metric experiments is that the Dice Coefficient will return a higher disambiguation accuracy than the other metrics because it incorporates the existence and non-existence of an element in the two vectors.

The purpose of the vector presentation experiments is to determine if the first-order vectors contain enough information to disambiguate the target word. In the NLM-WSD dataset, the context consists of the entire abstract in which the target word is used, which contains approximately 233.01 words for any given target word. Whereas the general English "Hard", "Line", "Serve", and "Interest" datasets contain approximately 36.79 words surrounding a target word, and the SENSEVAL-2 dataset contains approximately 107.11.

The hypothesis of this experiment is that given the size of the instances in the

dataset, first-order vectors should provide enough contextual information to disambiguate between the possible concepts.

First-order vectors contain highly frequent words that occur in the same window of context as the target word. The disadvantage of these vectors is that they are very sparse due to the limited number of words surrounding the target word. Second-order vectors attempt to alleviate the sparseness by including the features seen with the surrounding words themselves. Second-order vectors contain the words that occur *with* the words in the context surrounding the target word.

Table 6.2: Overall A-CUI Results

		Euclidean		Cosine		Dice	
Contextual Representation		o1	o2	o1	o2	o1	o2
Definition	CUI	0.50	0.40	0.53	0.50	0.52	0.54
	PAR	0.47	0.42	0.51	0.50	0.51	0.55
	CHD	0.45	0.46	0.49	0.45	0.48	0.51
	SIB	0.46	0.48	0.50	0.45	0.51	0.54
	SY	0.48	0.40	0.53	0.50	0.49	0.53
Terms	PT	0.43	0.40	0.53	0.44	0.40	0.50
	AT	0.46	0.39	0.49	0.51	0.46	0.53
Mapped Text	CUI 50	0.52	0.42	0.52	0.49	0.51	0.46
	CUI 100	0.43	0.43	0.51	0.44	0.47	0.43
	TERM 50	0.50	0.44	0.49	0.47	0.48	0.50
	TERM 100	0.49	0.41	0.49	0.43	0.49	0.47
Definition	Accuracy	0.47	0.43	0.51	0.48	0.50	0.53
Term	Accuracy	0.45	0.39	0.51	0.47	0.43	0.51
Mapped Text	Accuracy	0.49	0.43	0.50	0.46	0.49	0.47
Overall	Accuracy	0.47	0.42	0.51	0.47	0.47	0.50

Table 6.2 shows the overall accuracy of the first and second-order vectors when using the Euclidean distance, Cosine measure and Dice Coefficient using the following contextual representations extracted from the UMLS or MetaMap mapped text:

- UMLS CUI Definitions
  - the definition of a concepts CUI
  - the definition of the CUI + the parent definitions (PAR)
  - the definition of the CUI + the children definitions (CHD)
  - the definition of the CUI + the sibling definitions (SIB)

  - the definition of the CUI + the synonym definitions (SY)

- UMLS CUI Terms

  - a CUIs preferred terms (PT)
  - a CUIs associated terms (AT)

- MetaMap Mapped Text

  - 50 most frequent words in the same abstract as the possible concepts CUI (CUI 50)
  - 100 most frequent words in the same abstract as the possible concepts CUI (CUI 100)
  - 50 most frequent words in the same abstract as the terms associated with the possible concepts CUI (TERM 50)
  - 100 most frequent words in the same abstract as the terms associated with the possible concepts CUI (TERM 100)

The results show that first-order vectors obtain a higher overall disambiguation accuracy when using the Euclidean distance or Cosine Measure but second-order vectors obtain a higher disambiguation accuracy when using the Dice Coefficient. The results also show that the Cosine measure obtains a higher overall disambiguation accuracy than the Euclidean Distance or Dice Coefficient.

The results also show that the first-order Cosine results and the second-order Dice results obtain the highest two accuracies for all of the contextual representations. The overall accuracy of Definition experiments shows that the second-order Dice Coefficient (53%) obtains a higher disambiguation accuracy than the first-order Cosine Measure (51%), the overall accuracy of the Term experiments shows that the first-order Cosine and second-order Dice Coefficient results tied (51%), and the overall accuracy of MetaMapped Text experiments shows that the first-order Cosine Measure (50%) obtains a higher disambiguation accuracy than the second-order Dice Coefficient (47%).

The definitions do not represent the context in which a concept is used therefore the first-order vector of a definition is not actually a contextual representation of the concept. The second-order vectors contain the contextual representation of the words in the concept's definition, providing some contextual information, which may be why second-order vector obtain a higher disambiguation accuracy than first-order vectors for definitions.

The first-order concept vectors that use the words surrounding the concept in MetaMap mapped text actually provide a contextual representation of the concept indicating that first-order vectors provide enough information to disambiguate between the possible concepts.

6.2 UMLS CUI Definition Results

This section discusses the definition experiments. The purpose of these experiments is to determine if using a CUI definition provides enough information to distinguish between the possible concepts of an ambiguous word. The assumption is that the definition contains enough contextual information for disambiguation. [Patwardhan, 2003] show using the definitions from related concepts increased the results of their system that is designed to calculate the semantic relatedness between two concepts in WordNet.

The hypothesis of this experiment is that the context surrounding the words in the definition can be used to represent the context of the possible concept and adding information from related concepts will increase the disambiguation accuracy.

A-CUI uses the following context to create the concept vectors for this experiment:

- the definition of a concepts CUI
- the definition of the CUI + the parent definitions (PAR)
- the definition of the CUI + the children definitions (CHD)
- the definition of the CUI + the sibling definitions (SIB)
- the definition of the CUI + the synonym definitions (SY)

These experiments use the results obtained by A-CUI using second-order vectors and the Dice Coefficient. Table 6.3 shows the random and majority sense baselines and the accuracy for each of the above contexts. Table 6.4 shows the statistical significance between the results.

The results show that using just the CUI definition (CUI) obtains an accuracy of 54%. Adding the parent definitions (PAR) increases the accuracy (55%) by one percentage point, adding the children definitions (CHD) decreases the accuracy (51%) by three percentage points, adding the sibling definitions (SIB) has no effect on the overall accuracy and adding the source asserted synonym definition (SY) decreases the

Table 6.3: UMLS CUI Definition Results

	Baselines		Definitions				
Target Word	Rand.	Maj.	CUI	PAR	CHD	SIB	SY
adjustment	0.27	55.00	**0.69**	0.18	0.56	0.23	**0.69**
blood pressure	0.38	54.00	**0.53**	0.48	**0.53**	0.14	**0.53**
cold	0.14	49.00	0.01	0.15	0.05	**0.69**	0.01
condition	0.54	98.00	0.15	**0.71**	0.52	0.09	0.15
culture	0.44	89.00	0.11	0.10	**0.31**	0.09	0.11
degree	0.49	97.00	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
depression	0.46	100.00	**0.95**	0.92	**0.95**	**0.95**	**0.95**
determination	0.44	100.00	**1.00**	**1.00**	0.00	**1.00**	**1.00**
discharge	0.40	99.00	0.96	0.85	**0.97**	0.96	0.96
energy	0.44	99.00	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.52	50.00	0.50	**0.51**	0.50	0.50	0.50
extraction	0.43	94.00	0.05	0.05	0.05	**0.75**	0.05
failure	0.41	86.00	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**
fat	0.51	97.00	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
fit	0.56	100.00	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**
fluid	0.48	100.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
frequency	0.53	100.00	0.95	0.86	**0.97**	0.95	0.85
ganglion	0.52	93.00	**0.97**	0.95	**0.97**	0.34	**0.97**
glucose	0.54	91.00	**0.89**	0.87	0.55	0.87	**0.89**
growth	0.61	63.00	**0.37**	**0.37**	**0.37**	**0.37**	**0.37**
immunosuppression	0.48	59.00	0.54	0.55	0.52	**0.64**	0.54
implantation	0.49	83.00	**0.78**	0.30	0.63	0.76	**0.78**
inhibition	0.53	99.00	0.02	**0.69**	0.01	0.11	0.02
japanese	0.56	92.00	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	93.00	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
man	0.26	88.00	0.37	0.14	0.33	**0.45**	0.37
mole	0.39	99.00	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
mosaic	0.37	54.00	0.53	**0.55**	0.53	0.54	0.53
nutrition	0.42	51.00	0.19	0.22	0.18	**0.39**	0.19
pathology	0.45	86.00	0.25	**0.83**	0.71	0.18	0.25
pressure	0.28	100.00	**0.98**	**0.98**	0.97	0.97	**0.98**
radiation	0.52	61.00	**0.58**	**0.58**	0.57	0.57	**0.58**
reduction	0.36	82.00	**0.82**	**0.82**	**0.82**	**0.82**	**0.82**
repair	0.41	76.00	0.29	**0.35**	0.29	0.29	0.29
resistance	0.67	100.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
scale	0.32	100.00	**1.00**	0.98	0.02	**1.00**	**1.00**
secretion	0.53	99.00	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
sensitivity	0.31	96.00	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**
sex	0.29	80.00	0.16	0.16	**0.80**	0.19	0.16
single	0.53	99.00	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
strains	0.49	99.00	**0.14**	**0.14**	0.04	**0.14**	**0.14**
support	0.80	60.00	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**
surgery	0.50	98.00	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**
transient	0.52	99.00	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
transport	0.53	99.00	**0.98**	**0.98**	0.97	0.96	**0.98**
ultrasound	0.43	84.00	**0.73**	**0.73**	0.72	**0.73**	**0.73**
variation	0.54	80.00	0.20	0.20	**0.21**	0.20	0.20
weight	0.51	55.00	**0.57**	**0.57**	0.42	**0.57**	**0.57**
white	0.49	54.00	0.48	**0.59**	0.49	0.49	0.48
Overall Accuracy	0.46	83.86	**0.54**	0.55	0.51	0.54	0.53

Table 6.4: P-values using the Pairwise T-test for UMLS CUI Definition Results

	CUI	PAR	CHD	SIB	SY
Rand. baseline	0.07763	0.04134	0.15099	0.06678	0.08282
CUI		0.31972	0.24330	0.44434	0.16039

accuracy (53%) by one percentage point. There is no significant difference between these results. All of these results are higher than that of the random baseline but not the majority sense baseline which is to be expected because as previously stated the majority sense baseline is a supervised baseline. The difference in accuracy between the definitions and the random baseline is statistically significant except for using the CUI definition plus its children's definitions (CHD).

Table 6.5: Concepts in the NLM-WSD Dataset without a Definition

Definition	Target Word	CUI	Preferred Term
Parent (PAR)	blood pressure	C0428878	Arterial pressure
	determination	C0243075	adjudication
	failure	C0699796	failure
	growth	C0220844	growth
	japanese	C0022342	Japanese Population
	lead	C0373667	Lead measurement, quantitative
	mole	C0026386	Mole the mammal
	mole	C0349514	Benign melanocytic nevus of skin
	radiation	C0034618	Radiation therapy
	resistance	C0237834	resistance
	secretion	C0687157	Bodily secretions
	sensitivity	C0036667	Statistical sensitivity
	sex	C0036862	Coitus
	surgery	C0600001	Surgery specialty
	transient	C0040704	Transient Population Group
	transport	C0150390	Patient Transport
	ultrasound	C0041621	Ultrasonic Shockwave
Sibling (SIB)	association	C0699792	Relationship by association
	determination	C0680730	adjudication
	extraction	C0684295	extraction
	japanese	C0376247	Japanese language
	mosaic	C0700058	
	reduction	C0301630	Reduction (chemical
	resistance	C0683598	resistance
	sensitivity	C0312418	Personality Sensitivity
	single	C0087136	Unmarried
Synonymous (SY)	frequency	C0042023	Increased frequency of micturition

The results show that only the parent definitions provided enough contextual information to increase the overall accuracy of just using the concept definition. Adding the parent definitions decreases the number of possible concepts in the NLM-WSD dataset

that do not have a corresponding definition. There exist 113 possible concepts in the NLM-WSD dataset and of those concepts 49 of them do not have a corresponding definition as previously shown in Table 6.7. The addition of the parent definition provides a definition for 32 out of the 49 possible concepts who do not have one. Adding the SIB definitions only provides a definitions for nine out of the 49, and adding the SY definitions only provides a definition for one out of the 49. Table 6.5 lists these possible concepts.

A closer analysis between the CUI and PAR results with those obtained using the majority sense baseline indicates that A-CUI assigns a single concept to the test vectors a majority of the time. Table 6.6 shows the target word sorted based on their majority sense baseline as well as the number of possible concepts of a target word, the majority sense baseline, the CUI and PAR results and the difference between these results and the baseline at the individual target word level. The starred target words are those in which at least one of the possible concepts does not have a corresponding definition.

The difference scores for CUI show that 26 of the target words obtained less than a 10 percentage point difference and eight obtain a difference that is within 10 percentage points of the majority sense baseline. This indicates A-CUI assigns a single concept to the test vectors a majority of the time.

The reason is because some of the possible concepts for a target word do not have a corresponding definition in the UMLS so their vectors consists completely of zeros. The Dice Coefficient determines how close the test and concept vectors are by counting the number of times both vectors contain the same element and dividing it by the sum of the number of elements in the vectors. If there exists at least one element that the two vectors have in common it will assign the concept that has the definition. If this concept happens to also be the majority sense then the accuracy will be very high and otherwise the accuracy will be very low.

Table 6.7 shows a list of the CUIs in the NLM-WSD dataset that do not have a definition in the UMLS. 39 out of the 50 target words have a possible concept that does not have an associated definition in the UMLS which is 49 out of the 113 total possible concepts that exist in the dataset.

There are only twelve target words in the dataset in which all of the possible concept have a corresponding definition. These target words are shown in Table 6.8 with the

Table 6.6: Difference in Accuracy between the Baseline and Definition Results

	#	Maj.	CUI		PAR	
Target Word	Concepts	baseline	Accuracy	Difference	Accuracy	Difference
*cold	5	0.49	0.01	-0.48	0.15	-0.34
*evaluation	2	0.50	0.50	0.00	0.51	0.01
*nutrition	3	0.51	0.19	-0.32	0.22	-0.29
white	2	0.54	0.48	-0.06	0.59	0.05
*mosaic	3	0.54	0.53	-0.01	0.55	0.01
*blood pressure	3	0.54	0.53	-0.01	0.48	-0.06
weight	2	0.55	0.57	0.02	0.57	0.02
*adjustment	3	0.55	0.69	0.14	0.18	-0.37
immunosuppression	2	0.59	0.54	-0.05	0.55	-0.04
*support	2	0.60	0.80	0.20	0.80	0.20
*radiation	2	0.61	0.58	-0.03	0.58	-0.03
*growth	2	0.63	0.37	-0.26	0.37	-0.26
*repair	2	0.76	0.29	-0.47	0.35	-0.41
variation	2	0.80	0.20	-0.60	0.20	-0.60
*sex	3	0.80	0.16	-0.64	0.16	-0.64
*reduction	2	0.82	0.82	-0.00	0.82	-0.00
implantation	2	0.83	0.78	-0.05	0.30	-0.53
*ultrasound	2	0.84	0.73	-0.11	0.73	-0.11
*failure	2	0.86	0.83	-0.03	0.83	-0.03
pathology	2	0.86	0.25	-0.61	0.83	-0.03
man	3	0.88	0.37	-0.51	0.14	-0.74
culture	2	0.89	0.11	-0.78	0.10	-0.79
*glucose	2	0.91	0.89	-0.02	0.87	-0.04
*japanese	2	0.92	0.94	0.02	0.94	0.02
*lead	2	0.93	0.93	0.00	0.93	0.00
ganglion	2	0.93	0.97	0.04	0.95	0.02
*extraction	2	0.94	0.05	-0.89	0.05	-0.89
*sensitivity	3	0.96	0.02	-0.94	0.02	-0.94
*fat	2	0.97	0.93	-0.04	0.93	-0.04
*degree	2	0.97	0.03	-0.94	0.03	-0.94
*surgery	2	0.98	0.02	-0.96	0.02	-0.96
condition	2	0.98	0.15	-0.83	0.71	-0.27
discharge	2	0.99	0.96	-0.03	0.85	-0.14
*mole	3	0.99	0.99	-0.00	0.99	-0.00
*secretion	2	0.99	0.01	-0.98	0.01	-0.98
*energy	2	0.99	0.99	0.00	0.99	0.00
*transport	2	0.99	0.98	-0.01	0.98	-0.01
*transient	2	0.99	0.01	-0.98	0.01	-0.98
*strains	2	0.99	0.14	-0.85	0.14	-0.85
inhibition	2	0.99	0.02	-0.97	0.69	-0.30
*single	2	0.99	0.01	-0.98	0.01	-0.98
*pressure	3	1.00	0.98	-0.02	0.98	-0.02
*frequency	2	1.00	0.95	-0.05	0.86	-0.14
*fluid	2	1.00	1.00	0.00	1.00	0.00
*determination	2	1.00	1.00	0.00	1.00	0.00
*fit	2	1.00	0.06	-0.94	0.06	-0.94
*scale	3	1.00	1.00	0.00	0.98	-0.02
*depression	2	1.00	0.95	-0.05	0.92	-0.08
*resistance	2	1.00	1.00	0.00	1.00	0.00
Overall Accuracy		0.84	0.54		0.55	

Table 6.7: Possible Concepts in the NLM-WSD Dataset without Definitions

Target Word	CUI
adjustment	Individual Adjustment [C0376209]
association	Relationship by Association [C0699792]
blood pressure	Arterial Pressure [C0428878]
cold	Cold Sensation [C0234192]
degree	Degree [C0449286]
depression	Depression motion [C0460137]
determination	Adjudication [C0680730] Determination [C0243075]
energy	Vitality [C0424589]
evaluation	Health evaluation [C0175637]
extraction	Extraction [C0684295]
failure	Failure [C0699796]
fat	Obese Build [C0424612]
fit	Fit and Well [C0424576]
fluid	Liquid Substance, NOS [C0444611]
frequency	Increased Frequency of Micturition [C0042023]
glucose	Glucose Measurement [C0337438]
growth	Growth 2 [C0220844]
japanese	Japanese Language [C0376247] Japanese Population [C0022342]
lead	Lead Measurement, quantitative [C0373667]
mole	Mole the Mammal [C0026386] Benign Melanocytic Nevus of Skin [C0349514]
mosaic	Spatial Mosaic [C0439750] Mosaic [C0700058]
nutrition	Feeding and Dietary Regimes [C0600072]
pressure	Pressure - Action [C0460139] Baresthesia [C0234222]
radiation	Radiation Therapy [C0034618]
reduction	Reduction - Action [C0441610] Reduction Chemical [C0301630]
repair	Repair - Action [C0374711]
resistance	Resistance 1 [C0683598] Resistance 2 [C0237834]
scale	Integumentary scale [C0222045] Weight measurement scales [C0175659]
secretion	Bodily secretions [C0687157]
sensitivity	Statistical Sensitivity [C0036667] Personality Sensitivity [C0312418] Antimicrobial Susceptibility [C0427965]
sex	Coitus [C0036862]
single	Unmarried [C0087136]
strains	Muscle strain [C0080194]
strains	Microbiology subtype strains [C0456178]
support	support [C0183683]
surgery	Surgery specialty [C0600001]
transient	Transient Population Group [C0040704]
transport	Patient Transport [C0150390]
ultrasound	Ultrasonic Shockwave [C0041621]

accuracy of the random and majority sense baseline along with the accuracy of the definition results. A star next to the PAR, CHD, SIB and SY results indicates that the relation definition existed for at least one of the target possible concepts increasing the amount of contextual information.

Table 6.8: Results for Target Words with a UMLS CUI Definition

Target Word	Rand.	Maj.	CUI	PAR	CHD	SIB
condition	0.54	0.98	0.15	**0.71**	0.52	0.09
culture	0.44	0.89	0.11	0.10	**0.31**	0.09
discharge	0.40	0.99	0.96	0.85	**0.97**	0.96
ganglion	0.52	0.93	**0.97**	0.95	**0.97**	0.34
immunosuppression	0.48	0.59	0.54	0.55	0.52	**0.64**
implantation	0.49	0.83	**0.78**	0.30	0.63	0.76
inhibition	0.53	0.99	0.02	**0.69**	0.01	0.11
man	0.26	0.88	0.37	0.14	0.33	**0.45**
pathology	0.45	0.86	0.25	**0.83**	0.71	0.18
variation	0.54	0.80	0.20	0.20	**0.21**	0.20
weight	0.51	0.55	**0.57**	**0.57**	0.42	**0.57**
white	0.49	0.54	0.48	**0.59**	0.49	0.49
Overall Accuracy	0.47	**0.82**	0.45	0.54	0.51	0.41

These results tell a much different story. The overall accuracy for these target words when using just the concepts CUI definition is 45%. Adding the parent's definitions increases the accuracy by nine percentage points, adding the child definitions increases the accuracy by six percentage points while adding the sibling definitions decreases the accuracy by three percentage points. Each of the target words had a possible concept whose parent, child or sibling had a definition in the UMLS.

The overall conclusion of this experiment is that the addition of the parent and child definitions does increase the overall disambiguation accuracy although using the definitions in the UMLS in general is problematic because of the limited number of definitions that exist.

6.3 UMLS CUI Term Results

This section discusses the term experiments. The purpose of these experiments is to determine if using the terms used to describe a concept provides enough of context to

Table 6.9: UMLS CUI Term Results

	Baseline		Cosine		Dice	
Target Word	Rand.	Maj.	PT	AT	PT	AT
adjustment	0.27	0.55	0.37	0.48	**0.69**	0.68
blood pressure	0.38	0.54	0.42	0.27	**0.53**	0.02
cold	0.14	0.49	0.12	0.49	0.01	**0.86**
condition	0.54	0.98	0.21	**0.60**	0.15	0.15
culture	0.44	0.89	0.42	**0.54**	0.11	0.11
degree	0.49	0.97	0.06	**0.40**	0.03	0.03
depression	0.46	1.00	0.91	**0.95**	**0.95**	**0.95**
determination	0.44	1.00	0.97	0.54	**1.00**	0.00
discharge	0.40	0.99	0.92	0.49	**0.96**	0.88
energy	0.44	0.99	0.97	0.72	**0.99**	0.98
evaluation	0.52	0.50	0.44	0.48	0.50	**0.60**
extraction	0.43	0.94	0.10	**0.23**	0.05	0.05
failure	0.41	0.86	0.66	0.79	0.00	**0.83**
fat	0.51	0.97	0.85	0.79	**0.93**	0.58
fit	0.56	1.00	0.06	0.83	0.06	**1.00**
fluid	0.48	1.00	0.98	0.36	**1.00**	0.92
frequency	0.53	1.00	0.39	0.18	**0.95**	0.02
ganglion	0.52	0.93	0.93	0.93	**0.97**	0.93
glucose	0.54	0.91	0.83	0.45	**0.89**	0.85
growth	0.61	0.63	**0.46**	0.39	0.37	0.37
immunosuppression	0.48	0.59	**0.59**	0.52	0.54	0.53
implantation	0.49	0.83	0.61	0.60	**0.78**	0.70
inhibition	0.53	0.99	0.49	0.81	0.02	**0.99**
japanese	0.56	0.92	**0.94**	0.46	**0.94**	**0.94**
lead	0.21	0.93	0.86	0.83	**0.93**	0.07
man	0.26	0.88	**0.50**	0.46	0.37	0.00
mole	0.39	0.99	0.86	0.36	0.00	**0.99**
mosaic	0.37	0.54	**0.61**	0.43	0.53	0.49
nutrition	0.42	0.51	0.21	0.33	0.19	**0.39**
pathology	0.45	0.86	**0.31**	0.25	0.25	0.16
pressure	0.28	1.00	0.77	0.12	**0.98**	**0.98**
radiation	0.52	0.61	0.59	0.42	0.58	**0.67**
reduction	0.36	0.82	0.27	0.45	**0.82**	**0.82**
repair	0.41	0.76	0.34	**0.71**	0.29	0.29
resistance	0.67	1.00	0.00	0.33	**1.00**	**1.00**
scale	0.32	1.00	0.69	0.12	**1.00**	0.02
secretion	0.53	0.99	**0.39**	0.32	0.01	0.01
sensitivity	0.31	0.96	**0.24**	0.12	0.02	0.02
sex	0.29	0.80	0.24	0.49	0.16	**0.80**
single	0.53	0.99	**0.41**	0.31	0.01	0.08
strains	0.49	0.99	**0.73**	0.58	0.14	0.35
support	0.80	0.60	**0.80**	0.70	**0.80**	0.60
surgery	0.50	0.98	0.01	**0.07**	0.02	0.02
transient	0.52	0.99	0.33	0.50	0.01	**0.97**
transport	0.53	0.99	0.88	0.65	**0.98**	**0.98**
ultrasound	0.43	0.84	0.75	**0.84**	0.73	0.80
variation	0.54	0.80	**0.30**	0.27	0.20	0.20
weight	0.51	0.55	0.55	0.51	**0.57**	0.55
white	0.49	0.54	**0.59**	0.57	0.48	0.57
Overall Accuracy	0.46	0.83	**0.53**	0.49	0.50	**0.53**

distinguish between the possible concepts of the target word. [Pedersen et al., 2007] show using the words associated with a CUI in the Mayo Clinic Thesaurus provides enough contextual information to be used to determine the semantic relatedness between concepts in SNOMED-CT. The hypothesis of this experiment is that the term information in the UMLS will also provide enough information to distinguish between the possible concepts of a target word, and the addition of the associated terms will increase the accuracy of the results.

A-CUI uses the following context to create the concept vectors for this experiment:

- a CUIs preferred terms (PT)
- a CUIs associated terms (AT)

The preferred term is included in the list of associated terms. The experiments use the results obtained by A-CUI using first-order vectors (o1) and the Cosine measure as well as the results obtained by A-CUI using second-order vectors (o2) and the Dice Coefficient. Table 6.9 shows the random and majority sense baselines and the accuracy for each of the above contexts. Table 6.12 shows the statistical significance between the results.

The results show that using the preferred terms (PT) with the Cosine measure and first-order vectors obtains an accuracy of 53% while using the Dice Coefficient and second-order vectors obtains an accuracy of 50%. The results also show that using the associated terms (AT) with the Cosine measure and first-order vectors obtains an accuracy of 49% while using the Dice Coefficient and second-order vectors obtains an accuracy of 53%. The PT and AT results are higher than the random baseline but only the PT results are statistically significantly higher.

Not all of the concepts in the NLM-WSD dataset have an associated term. There are ten target words in the dataset in which one of the concepts does not have any associated terms and one target word in which neither of the concepts have any. Table 6.10 shows a list of these target words with the possible concepts.

The analysis of the majority sense baseline and the PT results using the second-order vectors and the Dice Coefficient indicate that A-CUI assigns a single concept to the test vectors a majority of the time. Table 6.11 shows the results of the majority concept baseline, these PT results using the Dice Coefficient and Cosine Measure and

Table 6.10: Target Words with No Associated Terms

Target Word	Possible Concept
blood pressure	Arterial pressure [C0428878]
determination	adjudication [C0243075]
failure	failure [C0699796]
japanese	Japanese Population [C0022342]
lead	Lead measurement, quantitative [C0373667]
mole	Mole the mammal [C0026386] Benign melanocytic nevus of skin [C0349514]
radiation	Radiation therapy [C0034618]
resistance	Resistance [C0237834]
secretion	Bodily secretions [C0687157]
sex	Coitus [C0036862]
surgery	Surgery specialty [C0600001]

the difference between these results and the baseline. There exists 90 target words whose difference is either under ten percentage points or over 90 when using the Dice Coefficient is 38 whereas only 19 when using the Cosine Measure.

Analysis of the preferred terms shows, for some of the target words, the preferred terms of their concepts are either almost identical or have overlapping words. For example, consider the target word *resistance* which has two possible concepts:

- Resistance 1 [C0683598]
- Resistance 2 [C0237834]

The preferred term for these two concepts is exactly the same except for a single digit to distinguish between them. There exists six other target words whose preferred terms are identical except for either a single digit or one of the terms ends in "NOS" such as "Support, NOS". These target words are: degree, extraction, failure, growth, pathology and support. 18 of the remaining 43 target words have overlapping words in their preferred terms. For example, the target word *depression* has two possible concepts in which the word *depression* occurs in both concepts:

- Mental Depression [C0011570]
- Depression motion [C0460137]

Table 6.11: Difference in Baseline and PT Results

	Maj.	Dice		Cosine	
Target Word	Baseline	Accuracy	Difference	Accuracy	Difference
cold	0.49	0.01	-0.48	0.12	-0.37
evaluation	0.50	0.50	0.00	0.44	-0.06
nutrition	0.51	0.19	-0.32	0.21	-0.30
white	0.54	0.48	-0.06	0.59	0.05
mosaic	0.54	0.53	-0.01	0.61	0.07
blood pressure	0.54	0.53	-0.01	0.42	-0.12
weight	0.55	0.57	0.02	0.55	-0.00
adjustment	0.55	0.69	0.14	0.37	-0.18
immunosuppression	0.59	0.54	-0.05	0.59	0.00
support	0.60	0.80	0.20	0.80	0.20
radiation	0.61	0.58	-0.03	0.59	-0.02
growth	0.63	0.37	-0.26	0.46	-0.17
repair	0.76	0.29	-0.47	0.34	-0.42
variation	0.80	0.20	-0.60	0.30	-0.50
sex	0.80	0.16	-0.64	0.24	-0.56
reduction	0.82	0.82	-0.00	0.27	-0.55
implantation	0.83	0.78	-0.05	0.61	-0.22
ultrasound	0.84	0.73	-0.11	0.75	-0.09
failure	0.86	0.00	-0.86	0.66	-0.20
pathology	0.86	0.25	-0.61	0.31	-0.55
man	0.88	0.37	-0.51	0.50	-0.38
culture	0.89	0.11	-0.78	0.42	-0.47
glucose	0.91	0.89	-0.02	0.83	-0.08
japanese	0.92	0.94	0.02	0.94	0.02
lead	0.93	0.93	0.00	0.86	-0.07
ganglion	0.93	0.97	0.04	0.93	0.00
extraction	0.94	0.05	-0.89	0.10	-0.84
sensitivity	0.96	0.02	-0.94	0.24	-0.72
fat	0.97	0.93	-0.04	0.85	-0.12
degree	0.97	0.03	-0.94	0.06	-0.91
surgery	0.98	0.02	-0.96	0.01	-0.97
condition	0.98	0.15	-0.83	0.21	-0.77
discharge	0.99	0.96	-0.03	0.92	-0.07
mole	0.99	0.00	-0.99	0.86	-0.13
secretion	0.99	0.01	-0.98	0.39	-0.60
energy	0.99	0.99	0.00	0.97	-0.02
transport	0.99	0.98	-0.01	0.88	-0.11
transient	0.99	0.01	-0.98	0.33	-0.66
strains	0.99	0.14	-0.85	0.73	-0.26
inhibition	0.99	0.02	-0.97	0.49	-0.50
single	0.99	0.01	-0.98	0.41	-0.58
pressure	1.00	0.98	-0.02	0.77	-0.23
frequency	1.00	0.95	-0.05	0.39	-0.61
fluid	1.00	1.00	0.00	0.98	-0.02
determination	1.00	1.00	0.00	0.97	-0.03
fit	1.00	0.06	-0.94	0.06	-0.94
scale	1.00	1.00	0.00	0.69	-0.31
depression	1.00	0.95	-0.05	0.91	-0.09
resistance	1.00	1.00	0.00	0.00	-1.00
Overall Accuracy	0.84	0.50		0.53	

Table 6.12: P-values using the Pairwise T-test for UMLS CUI Term Results

		Random	PT		AT
		Baseline	Cosine (o1)	Dice (o2)	Cosine (o1)
PT	Cosine (o1)	0.05129			
	Dice (o2)	0.21739	0.24691		
AT	Cosine (o1)	0.14876	0.17520	0.43808	
	Dice (o2)	0.09748	0.47442	0.34014	0.23131

There is a star next to these 15 words in Table 6.11. Appendix E contains a complete list of all of the target words in the NLM-WSD dataset and the preferred terms of their concepts.

The analysis also shows that the number of words in the concept vector is on average 1.80 words therefore the number of non-zero elements in the concept vector is going to be on average 1.80. This indicates that the Cosine measure is not as greatly affected by the small number of elements in the concept vectors. The addition of the associated terms brings the average up to 2.42, which decreases the overall accuracy when using the Cosine measure.

The overall conclusions of this experiment is that the preferred terms are distinct enough to provide a distinction between the concepts, but adding the associated terms does not provide any more information to aid in the disambiguation process.

6.4 MetaMap Mapped Text Results

This section discusses the MetaMap mapped text experiments. The purpose of these experiments is to determine if using highly frequent words that exist in the same abstract as the CUI assigned by MetaMap or its associated terms provide enough unique contextual information in order to distinguish between the possible concepts of the target word. [Pedersen et al., 2007] show using the words associated with a CUI in an outside source provides enough contextual information to be used to determine the semantic relatedness between concepts in SNOMED-CT. The hypothesis of this experiment is that the additional information will provide a better contextual representation for the possible concepts and increasing the accuracy of the results.

A-CUI uses the following context to create the concept vectors for this experiment:

Table 6.13: MetaMap Mapped Text Results

		Baseline					
Target Word	#	Random	Majority	CUI 50	CUI 100	TERM 50	TERM 100
adjustment	3	0.27	55.00	0.31	0.23	0.41	**0.51**
blood pressure	3	0.38	54.00	**0.42**	0.32	0.26	0.19
cold	5	0.14	49.00	**0.44**	0.43	0.16	0.03
condition	2	0.54	98.00	0.86	**0.95**	0.74	0.86
culture	2	0.44	89.00	0.51	**0.54**	0.26	0.14
degree	2	0.49	97.00	0.28	0.20	**0.29**	0.20
depression	2	0.46	100.00	**0.41**	0.15	0.06	0.11
determination	2	0.44	100.00	0.27	0.32	**0.43**	0.39
discharge	2	0.40	99.00	**0.75**	0.59	0.53	0.49
energy	2	0.44	99.00	0.61	0.68	0.88	**0.91**
evaluation	2	0.52	50.00	**0.59**	0.53	0.45	0.51
extraction	2	0.43	94.00	0.60	**0.66**	0.58	0.50
failure	2	0.41	86.00	**0.69**	**0.69**	0.66	0.66
fat	2	0.51	97.00	0.26	**0.41**	0.25	0.16
fit	2	0.56	100.00	0.78	0.56	**0.83**	0.50
fluid	2	0.48	100.00	0.60	**0.66**	0.34	0.17
frequency	2	0.53	100.00	0.04	0.02	**0.74**	0.70
ganglion	2	0.52	93.00	**0.93**	**0.93**	**0.93**	**0.93**
glucose	2	0.54	91.00	0.64	**0.70**	0.40	0.53
growth	2	0.61	63.00	0.61	0.61	0.62	**0.63**
immunosuppression	2	0.48	59.00	0.52	0.41	0.63	**0.70**
implantation	2	0.49	83.00	**0.54**	0.42	0.39	0.32
inhibition	2	0.53	99.00	0.76	**0.83**	0.62	0.73
japanese	2	0.56	92.00	0.81	0.84	**0.94**	**0.94**
lead	2	0.21	93.00	**0.93**	**0.93**	0.90	**0.93**
man	3	0.26	88.00	0.42	0.46	0.28	**0.47**
mole	3	0.39	99.00	0.27	0.10	0.79	**0.92**
mosaic	3	0.37	54.00	0.35	0.37	0.38	**0.42**
nutrition	3	0.42	51.00	**0.39**	0.36	0.27	0.35
pathology	2	0.45	86.00	0.55	0.55	**0.66**	0.60
pressure	3	0.28	100.00	0.46	0.39	**0.61**	0.39
radiation	2	0.52	61.00	0.53	0.52	0.52	**0.54**
reduction	2	0.36	82.00	0.18	0.18	**0.73**	**0.73**
repair	2	0.41	76.00	**0.65**	0.53	0.56	0.54
resistance	2	0.67	100.00	**1.00**	**1.00**	0.67	0.33
scale	3	0.32	100.00	0.62	0.65	**0.74**	0.68
secretion	2	0.53	99.00	0.23	**0.29**	**0.29**	0.27
sensitivity	3	0.31	96.00	**0.69**	0.67	0.47	0.45
sex	3	0.29	80.00	0.51	0.41	**0.65**	0.58
single	2	0.53	99.00	**0.40**	0.37	0.07	0.08
strains	2	0.49	99.00	**0.81**	0.78	0.34	0.35
support	2	0.80	60.00	0.20	**0.40**	0.30	0.20
surgery	2	0.50	98.00	0.02	**0.03**	0.02	**0.03**
transient	2	0.52	99.00	0.59	**0.61**	0.40	0.40
transport	2	0.53	99.00	0.20	0.32	0.78	**0.85**
ultrasound	2	0.43	84.00	**0.76**	0.74	0.67	0.74
variation	2	0.54	80.00	**0.77**	0.75	0.32	0.43
weight	2	0.51	55.00	0.45	0.49	0.49	**0.57**
white	2	0.49	54.00	0.39	**0.47**	0.43	0.43
Overall Accuracy	2.26	0.46	**83.86**	0.52	0.51	0.50	0.49

- 50 most frequent words in the same abstract as the possible concepts CUI (CUI 50)
- 100 most frequent words in the same abstract as the possible concepts CUI (CUI 100)
- 50 most frequent words in the same abstract as the terms associated with the possible concepts CUI (TERM 50)
- 100 most frequent words in the same abstract as the terms associated with the possible concepts CUI (TERM 100)

These experiments use the results obtained by A-CUI using first-order vectors and the Cosine measure. Table 6.13 shows the random and majority sense baselines and the accuracy for each of the above contexts. Table 6.14 shows the statistical significance between the results.

Table 6.14: P-values using the Pairwise T-test for the MetaMap Results

	CUI 50	CUI 100	TERM 50	TERM 100
Random Baseline	0.03736	0.06286	0.09790	0.18972
CUI 50		0.19162	0.31225	0.22059
CUI 100			0.43254	0.31268
TERM 50				0.19544

The results show that using the top 50 most frequent words surrounding the CUI (CUI 50) obtains an accuracy of 52% and using the top 100 most frequent words (CUI 100) obtains an accuracy of 51%. Using the top 50 most frequent words surrounding the terms (TERM 50) associated with the CUI obtains an accuracy of 50%, and using the top 100 most frequent words (TERM 100) obtains an accuracy of 49%. In each case, adding the addition 50 words to the context reduces the overall accuracy by one percentage point. The difference in the results is not statistically significant. The results also show that the overall accuracy is higher than the random baseline for each of the contextual representations, but only the CUI 50 results are statistically significantly higher.

Analysis of the concept vectors associated with each of the possible concepts of the target word show that in these cases there is considerable amount of overlap between the words in the context. Table 6.16 shows the overlap of words in the contexts of each

Table 6.15: Analysis of the Target Word *fat*

Word	C0424612	C0015677	Word	C0424612	C0015677	Word	C0424612	C0015677
abdominal	x		acid	x	x	body	x	x
both	x	x	carbohydrate	x	x	cholesterol	x	x
compared	x	x	content	x	x	diet	x	x
dietary	x	x	diets	x	x	during	x	x
effect	x	x	effects	x	x	energy	x	x
fat	x	x	fats		x	fatty	x	x
fed	x	x	group	x	x	high	x	x
high-fat	x	x	higher	x	x	increase	x	x
increased	x	x	insulin	x	x	intake	x	x
lean	x		less	x	x	levels	x	x
liver	x	x	low	x	x	lower	x	x
mass	x	x	milk	x	x	muscle	x	x
oil		x	patients	x	x	percentage	x	x
plasma	x	x	protein	x	x	rats	x	x
significant	x	x	significantly	x	x	study	x	x
subjects	x	x	these	x	x	tissue	x	x
total	x	x	weight	x	x	p	x	x
+/-	x	x						

of the possible concepts of a target word. For example, consider the target word fat which has two possible concepts: Obese Build [C0424612] and Fatty acid glycerol esters [C0015677]. Table 6.15 shows the top 50 most frequent words that exist in the same abstract as the CUIs where 48 of the words exist in both contexts.

Nine target words in the dataset contain concepts whose CUI did not exist in the MetaMapped 2005 Medline baseline:

- determination
- failure
- lead
- radiation
- resistance
- secretion
- surgery
- variation

Table 6.16: Overlap of Words Between the Context of the Possible Concepts

Target Word	CUI 50	CUI 100	TERM 50	TERM 100
adjustment	49/55	98/108	42/106	88/205
blood pressure	48/91	96/175	1/99	7/193
cold	51/129	113/234	62/168	116/328
condition	0/50	0/100	13/87	36/164
culture	46/54	92/108	7/93	30/170
degree	50/50	100/100	46/54	96/104
depression	48/52	99/101	0/101	0/201
determination	0/50	0/100	0/50	0/100
discharge	38/62	83/117	0/100	11/189
energy	49/51	99/101	3/97	4/196
evaluation	19/81	38/162	7/93	22/178
extraction	50/50	100/100	49/51	96/104
failure	0/0	0/0	0/50	0/100
fat	48/52	99/101	10/90	22/178
fit	28/72	63/137	7/93	25/175
fluid	43/57	87/113	16/84	40/160
frequency	0/50	0/100	2/98	18/182
ganglion	2/98	14/186	4/96	12/188
glucose	14/86	39/161	17/83	52/148
growth	48/52	99/101	50/50	100/100
immunosuppression	44/56	88/112	4/96	13/187
implantation	25/75	67/133	8/92	16/184
inhibition	50/50	100/100	0/101	1/199
japanese	7/93	19/181	0/0	0/0
lead	0/50	0/100	0/50	0/100
man	37/100	76/195	36/109	74/208
mole	49/99	98/199	0/50	0/100
mosaic	50/78	100/143	50/90	100/171
nutrition	50/74	100/142	14/86	48/152
pathology	47/53	96/104	46/54	92/108
pressure	50/52	99/104	49/68	96/129
radiation	11/89	26/174	0/50	0/100
reduction	0/50	0/100	13/87	35/165
repair	37/63	75/125	0/100	2/198
resistance	0/50	0/100	43/57	92/108
scale	50/55	100/104	50/63	100/130
secretion	0/50	0/100	0/50	0/100
sensitivity	50/52	100/104	52/61	104/116
sex	44/56	96/104	19/81	43/157
single	49/51	99/101	1/99	8/192
strains	50/50	100/100	0/50	0/100
support	47/53	97/103	4/96	15/185
surgery	0/50	0/100	0/50	0/100
transient	43/57	88/112	8/92	19/181
transport	50/50	99/101	7/93	15/185
ultrasound	3/97	8/192	0/50	0/100
variation	0/50	0/100	47/53	98/102
weight	45/55	90/110	21/79	44/156
white	48/52	99/101	12/88	33/167

The reason is because the CUIs in the NLM-WSD dataset come from the 1999 tag set whereas the CUIs mapped to the 2005 Medline baseline come from the 2005 tag set. Given the nature of the UMLS the CUIs have changed over time.

The results also show that CUI 50 and CUI 100 have five target words have possible concepts that contain the same context:

- degree
- extraction
- inhibition
- strains
- transport

The reason is because MetaMap maps terms, not words, to CUIs based on the centrality, variation, coverage and cohesiveness of the term. Consider the single word term *transport*, MetaMap maps this term to the two possible CUIs in the UMLS: Biological Transport [C0005528] and Patient Transport [C0150390]. Now consider the multi-word term *patient transport*, MetaMap maps this term to only Patient Transport [C0150390] because of the coverage and cohesiveness of the term compared to the preferred term of the CUI. This happens enough in the 2005 Medline baseline such that only five out of the 40 target words whose possible concepts all were found in the baseline extracted exactly the same context.

There exists ten target words that do not have any associated terms in the UMLS, seen previously in Table 6.10, and three target words whose associated terms do not exist in the 2005 Medline baseline, shown in Table 6.17. There also exists one target word, *growth*, that contains the same context for each of its target words. This indicates that the words obtained surrounding the associated terms is more distinct than using the words surrounding the CUI.

The overall conclusions of this experiment is that the words that occur frequently in the same abstract as the CUI or associated terms provide enough distinct information in order to disambiguate between the concepts. Surprisingly, using the associated terms rather than the CUI did not increase the accuracy of the overall results indicating that MetaMap is able to accurately identify the appropriate concept of an ambiguous word based on its term information (not the context that it was used) well enough to provide distinct contextual information about that term.

Table 6.17: Terms Not in 2005 Medline Baseline

Target Word	CUI	Associated Terms
japanese	Japanese Language [C0376247]	idioma japon japanese language japanese language idioma japones
	Japanese Population [C0022342]	No terms
strains	Microbiology Subtype Strains [C0456178]	microbiological strain microbiological strains
ultrasound	Ultrasonic Shockwave [C0041621]	shock waves, ultrasonic shockwaves, ultrasonic

6.5 Previous Work Experiments

This section discusses the previous work experiments. There has been very little previous work in methods that do not use supervised learning algorithms to disambiguate words in biomedical text. [Humphrey et al., 2006] introduce a knowledge-based method that determines the appropriate semantic type of a target word with the assumption that the possible concepts of the target word have a unique semantic type. For example, consider the target word *culture* that has two possible concepts, Anthropological Culture [C0010453] and Laboratory Culture [C0430400], each with a different semantic type. The semantic type for Anthropological Culture [C0010453] is "Idea or Concept" while the semantic type for Laboratory Culture [C0430400] is "Laboratory Procedure". This method first assigns the instance containing the target word *culture* one of the two semantic types and then it determines the appropriate concept based on the assigned semantic type.

Identifying the semantic type of a target word is also a simpler problem than identifying its concept because semantic types are a coarser grained categorization.

[Humphrey et al., 2006] evaluate their method using 45 out of the 50 target words in the NLM-WSD dataset. The target word *association* was removed because all of the instances in that set were tagged as "None". The target words *cold*, *man*, *sex* and *weight* were removed because the possible concepts of the target have the same semantic type.

Table 6.18 shows the results of the random and majority sense baseline, the results

reported by [Humphrey et al., 2006], the definition results (CUI and PAR) discussed previously in Section 6.2 and the MetaMap mapped text results (CUI 50 and TERM 50) discussed in Section 6.4.

The results show the method proposed by [Humphrey et al., 2006] obtains an overall accuracy of 75% on the Humphrey subset while the CUI and PAR results obtain an accuracy of 56% and 58%, and the CUI 50 and TERM 50 obtain an accuracy of 53% and 50%. Only the PAR results obtain an statistically significantly higher accuracy than the random baseline for this subset as shown in Table 6.19.

Although the method proposed by [Humphrey et al., 2006] obtains a significantly higher disambiguation accuracy than the results obtained by A-CUI, the disadvantage of this method is that if two possible concepts have the same semantic type(s) the system would not be able to disambiguate between them. For example, consider the target word *man* which has three possible concepts:

- Male [C0024554]
- Man [C0025266]
- Homo sapiens [C0086418]

The concepts Man [C0025266] and Homo sapiens [C0086418] both have the semantic type "Population Group". Now consider the sentence:

Man has existed on this planet for roughly 50000 years.

The *man* in this sentence is not referring to the population group consisting of men but the group consisting of homo sapiens.

6.6 Conclusions

In this chapter, four sets of experiments were conducted. The first experiments investigated the metric and contextual representation options. The overall conclusion of this experiment is that the vector type and metric are dependent on the type of information being used as the context of the possible concept. The results showed, when using the words surrounding a CUI in the MetaMapped 2005 Medline baseline as the context for a concept, the Cosine measure obtained the highest overall disambiguation accuracy,

Table 6.18: Overall Results of A-CUI and Related Work

	Baseline		Humphrey	Definition		MetaMapped Text	
Target Word	Rand.	Maj.	et. al. 2006	CUI	PAR	CUI 50	TERM 50
adjustment	0.27	0.55	**0.77**	0.69	0.18	0.31	0.41
blood pressure	0.38	0.54	0.42	**0.53**	0.48	0.42	0.26
cold	0.14	0.49		0.01	0.15	**0.44**	0.16
condition	0.54	0.98	**0.93**	0.15	0.71	0.86	0.74
culture	0.44	0.89	**1.00**	0.11	0.10	0.51	0.26
degree	0.49	0.97	**0.98**	0.03	0.03	0.28	0.29
depression	0.46	1.00	0.95	**0.95**	0.92	0.41	0.06
determination	0.44	**1.00**	**1.00**	**1.00**	**1.00**	0.27	0.43
discharge	0.40	0.99	0.93	**0.96**	0.85	0.75	0.53
energy	0.44	**0.99**	0.70	**0.99**	**0.99**	0.61	0.88
evaluation	0.52	0.50	**0.60**	0.50	0.51	0.59	0.45
extraction	0.43	0.94	**0.98**	0.05	0.05	0.60	0.58
failure	0.41	0.86	**0.94**	0.83	0.83	0.69	0.66
fat	0.51	0.97	0.75	**0.93**	**0.93**	0.26	0.25
fit	0.56	**1.00**	**1.00**	0.06	0.06	0.78	0.83
fluid	0.48	**1.00**	0.06	**1.00**	**1.00**	0.60	0.34
frequency	0.53	1.00	0.90	**0.95**	0.86	0.04	0.74
ganglion	0.52	0.93	0.94	**0.97**	0.95	0.93	0.93
glucose	0.54	0.91	0.39	**0.89**	0.87	0.64	0.40
growth	0.61	0.63	**0.70**	0.37	0.37	0.61	0.62
immunosuppression	0.48	0.59	**0.75**	0.54	0.55	0.52	0.63
implantation	0.49	0.83	**0.94**	0.78	0.30	0.54	0.39
inhibition	0.53	**0.99**	**0.99**	0.02	0.69	0.76	0.62
japanese	0.56	0.92	0.55	**0.94**	**0.94**	0.81	**0.94**
lead	0.21	0.93	0.39	**0.93**	**0.93**	**0.93**	0.90
man	0.26	0.88		0.37	0.14	**0.42**	0.28
mole	0.39	0.99	0.98	**0.99**	**0.99**	0.27	0.79
mosaic	0.37	0.54	**0.68**	0.53	0.55	0.35	0.38
nutrition	0.42	0.51	0.35	0.19	0.22	**0.39**	0.27
pathology	0.45	0.86	0.75	0.25	**0.83**	0.55	0.66
pressure	0.28	1.00	0.12	**0.98**	**0.98**	0.46	0.61
radiation	0.52	0.61	**0.79**	0.58	0.58	0.53	0.52
reduction	0.36	0.82	**1.00**	0.82	0.82	0.18	0.73
repair	0.41	0.76	**0.86**	0.29	0.35	0.65	0.56
resistance	0.67	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.67
scale	0.32	**1.00**	0.60	**1.00**	0.98	0.62	0.74
secretion	0.53	0.99	**0.94**	0.01	0.01	0.23	0.29
sensitivity	0.31	0.96	**0.83**	0.02	0.02	0.69	0.47
sex	0.29	0.80		0.16	0.16	0.51	**0.65**
single	0.53	0.99	**1.00**	0.01	0.01	0.40	0.07
strains	0.49	0.99	**0.98**	0.14	0.14	0.81	0.34
support	0.80	0.60	**1.00**	0.80	0.80	0.20	0.30
surgery	0.50	0.98	**0.93**	0.02	0.02	0.02	0.02
transient	0.52	**0.99**	**0.99**	0.01	0.01	0.59	0.40
transport	0.53	0.99	**0.98**	0.98	0.98	0.20	0.78
ultrasound	0.43	0.84	**0.81**	0.73	0.73	0.76	0.67
variation	0.54	0.80	0.73	0.20	0.20	**0.77**	0.32
weight	0.51	0.55		**0.57**	**0.57**	0.45	0.49
white	0.49	0.54	0.55	0.48	**0.59**	0.39	0.43
Overall Accuracy	0.46	0.84		0.54	0.55	0.52	0.50
Humphrey subset	0.57	0.86	0.75	0.56	0.58	0.53	0.50

Table 6.19: P-values using the Pairwise T-test for A-CUI and Related Work

	Humphrey	Definition		Mapped Text	
	et. al. 2006	CUI	PAR	CUI 50	TERM 50
Random Baseline	0.000001	0.07763	0.04134	0.03736	0.09790
Humphrey et. al. 2006		0.00764	0.00935	0.00052	0.00011

and when using the CUI definitions as a concepts context the Dice Coefficient obtained the highest accuracy.

The results also showed that the first-order vectors obtained the highest overall disambiguation accuracy when using the words in the 2005 Medline baseline as the concepts contextual representation but second-order vectors obtained the highest overall accuracy when using the CUI definitions. This indicates that the definitions alone did not provide enough disambiguation information requiring the second-order information to determine the correct concept of a target word.

The remaining three experiments investigated using the following context information extracted from the UMLS and MetaMap mapped text:

- the definition of a concepts CUI (CUI)
- the definition of the CUI + the parent definitions (PAR)
- the definition of the CUI + the children definitions (CHD)
- the definition of the CUI + the sibling definitions (SIB)
- the definition of the CUI + the synonym definitions (SY)
- a CUIs preferred terms (PT)
- a CUIs associated terms (AT)
- 50 most frequent words in the in the same abstract as the possible concepts CUI (CUI 50)
- 100 most frequent words in the same abstract as the possible concepts CUI (CUI 100)
- 50 most frequent words in the same abstract as the terms associated with the possible concepts CUI (TERM 50)
- 100 most frequent words in the same abstract as the terms associated with the possible concepts CUI (TERM 100)

The overall results showed that using the CUI definitions obtained the highest overall accuracy. The problem though is that only a limited number of definitions actually exist in the UMLS. To use this method effectively another source of definitions would have to be found to supplement the CUI definitions. The results showed that using the PAR definitions in conjunction with the CUI's definition increased the overall accuracy but the parent definitions have the same limitation, only a limited number of them exist.

The context using the words surrounding the CUI in the MetaMapped 2005 Medline baseline shows the most promise even though the results were not as high as the definition. The CUIs in the NLM-WSD dataset come from the 1999 tag set while the CUIs in the 2005 Medline baseline come from the 2005 tag set therefore not all the CUIs in the NLM-WSD data exist in this baseline. Given a text in which the concept and the data are tagged with the same tag set, the results have the potential to increase.

Lastly, a comparative analysis between A-CUI and the knowledge-based method proposed by [Humphrey et al., 2006] was conducted. The results showed that their method obtained a higher overall disambiguation accuracy, but their system performs a simpler task by determining the appropriate semantic type of a target word with the assumption that the possible concepts of the target word have a unique semantic type. The disadvantage of this method is that if two possible concepts have the same semantic type the system would not be able to disambiguate between them.

The overall conclusion of the experiments conducted in this chapter is that CUI information extracted from the UMLS or MetaMap mapped text provides a unique contextual representation about the possible concepts of a target words in order to disambiguate between them. A more detailed analysis of these results is in Chapter 9.

Chapter 7

Related Work

This chapter discusses previous work related to this dissertation. The work included in this chapter has a direct relation to work that has been conducted in the biomedical domain. Section 7.1 discusses the biomedical and general English features that have been used to disambiguate words in both biomedical and general English text. Section 7.2 discusses the methods that have been used in general English that have been applied to the biomedical domain as well as those created specifically for the biomedical domain. These methods are classified into three categories: supervised methods, clustering methods, and knowledge-based methods. A general description of each of these methods is described in Chapter 2.

7.1 WSD Features

There are a number of different features that have been used in WSD. This section classifies them into two categories: biomedical and general English features. The general English features included in this section are those that were originally created to disambiguate words English text and later applied to biomedical text.

General English features have been shown to perform quite well when disambiguating words in biomedical text. Recently though, there has been work using biomedical features to disambiguate words in biomedical text such as the methods proposed by [Leroy and Rindflesch, 2004], [Humphrey et al., 2006], [Fan and Friedman, 2008] and

[Stevenson et al., 2008]. These methods use domain knowledge to help distinguish between the concepts of a target word. The following subsection discuss these two types of features in more detail.

7.1.1 General English Features

This section splits the general English features into three categories:

- lexical features
- syntactic features
- semantic features

Lexical features are features that are obtained by analyzing the target word and its surrounding words such as bag-of-words, bigrams, and collocations. Syntactic features are features that are obtained by analyzing the syntactic structure of the context containing the target word such as a word's morphology or part-of-speech (POS). Semantic features are features that are extracted from a knowledge-source such as the semantic similarity between two words, or classification. The remainder of this section discusses each of these types of features in more detail.

Lexical Features

This subsection discusses the lexical features shown in Table 7.1. The related work in this table is categorized into two section, those that used the features to disambiguate words in general English and those that used them to disambiguate words in biomedical text.

The *bag-of-words* feature consists of the words surrounding the target word without respect to order. [Gale et al., 1992] call this approach the "Information Retrieval Approach to Sense Disambiguation" because this feature is commonly used in information retrieval where documents are treated as a bag-of-words where word order is ignored and the words are considered independent of each other. The authors introduce using the bag-of-words feature for the task of WSD with the assumption that the words surrounding the target word are going to be different for each of the possible concepts. [Leacock et al., 1993], [Mooney, 1996] and [Pedersen, 2000] use the bag-of-words feature in their supervised method to disambiguate words in general English.

Table 7.1: Lexical WSD Features

General English	Bag-of-words	Unigrams	Bigrams	Co-occurrences	Collocations
[Gale et al., 1992]	x				
[Leacock et al., 1993]	x				
[Mooney, 1996]	x				
[Pedersen, 2000]	x				
[Ng and Lee, 1996]	x				x
[Lee and Ng, 2002]	x				x
[Patwardhan, 2003]		x			
[Mohammad and Hirst, 2006]		x			
[Mohammad and Pedersen, 2004]		x	x		
[Purandare and Pedersen, 2004]			x	x	
[Schütze, 1998]				x	
[Pedersen and Bruce, 1997]				x	x
[Pedersen and Bruce, 1998]				x	x
[Hirst, 1987]					x
Biomedical	**Bag-of-words**	**Unigrams**	**Bigrams**	**Co-occurrences**	**Collocations**
[Joshi et al., 2005]		x	x		
[Liu et al., 2004]	x			x	x
[Stevenson et al., 2008]		x	x		x
[Savova et al., 2008]	x				
[Alexopoulou et al., 2009]				x	

The term *unigrams* has been used to describe the "words" in the *bag-of-words* such as in the method proposed by [Yarowksy and Florian, 2003]. The unigram feature is part of a larger set of features called *ngrams*. Ngrams are defined as an ordered set of n words that occur frequently together. For example, the term *unigram* refers to a single word and *bigrams* refers to an ordered set of two words. The ngram feature can be thought of as an extension of the bag-of-words. Once the ngrams are obtained the features themselves are independent of each other. An ngram feature is selected based on the number of times it occurs in a corpus or a measure of association between the words in the ngram. [Pedersen, 2001] and [Mohammad and Pedersen, 2004] investigate using the Log Likelihood Ratio in their methods that disambiguate words in general English. [Stevenson et al., 2008] investigate using this feature to disambiguate words in biomedical text.

The assumption behind using ngrams is that the larger n the less ambiguous the

ngram; bigrams contain less ambiguity than unigrams and trigrams contain less ambiguity than bigrams. There is a disadvantage of using too large of a value for n though because the larger n, the less likely the ngram will occur. [Mohammad and Pedersen, 2004] compare unigrams and bigrams using their proposed supervised method. They evaluate the features on general English text and report that there was no significant difference in using one over the other. This is counter to the results shown by [Joshi et al., 2005], who found that using unigrams obtained a higher accuracy than bigrams when evaluated on biomedical text.

[Liu et al., 2004] extend the ngram feature to include the orientation of the ngram with respect to the target word. The orientation refers to whether the ngram is to the right or left of the target word which allows for the same word to be used as two different features depending on where it exists in the instance. The assumption is that with fine grained concepts the location of the surrounding word might be different depending on the possible concept.

A co-occurrence is a word or set of words that occur frequently with the target word and the order in which the words occur does not matter. For example, *interest payment* and *payment interest* are considered a single feature. [Pedersen and Bruce, 1997] and [Pedersen and Bruce, 1998] introduce a clustering method that incorporates frequently co-occurring pairs of words. [Purandare and Pedersen, 2004] compare both co-occurrences, unigrams, and bigrams in their clustering method. They report that for very sparse data, co-occurrences obtained the highest disambiguation, otherwise bigrams obtained the highest accuracy.

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. [Weaver, 1949] asked the question:

"If one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word ... The practical question is: What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"

Various researchers have investigated using different size windows in their methods in

order to answer this question. [Choueka and Lusignan, 1985] conducted an experiment to determine what size window is needed for humans to determine the appropriate concept of a target word. The authors report that only a small window size of two or three is needed indicating that the words closest to the target word matter the most in disambiguation.

This experiment lead to the introduction of *collocations*, by [Hirst, 1987], which are an ordered set of words that include the target word and occur frequently together. Since then there have been a variety of different types of features that include collocations. [Liu et al., 2004] include two word collocations as features, and [Ng and Lee, 1996] extract the collocations with a left offset of negative two and a right offset of one of the target word. For example, taking the target word *interest*, a possible collocation would be "in the interest of". [Pedersen and Bruce, 1998] propose restricting the collocation to include only content words to the left and right of the target word referring to them as *content collocations* and the former as *unrestricted collocations*. Stevenson, et. al. use what they describe as *local collocations* which they define as:

- bigrams and trigrams containing the target word
- preceding/following content words in the same sentence as the target word.

Although, [Choueka and Lusignan, 1985] showed that only a small window size is required for humans to disambiguate a word, [Gale et al., 1992] showed that larger the window size was needed for the computer. This was also shown by [Joshi et al., 2005] in the biomedical domain who report using ngrams in the same abstract as the target word obtains a higher accuracy than using ngrams in the same sentence as the target word. [Liu et al., 2004] compared bag-of-words and collocations reporting that the collocations did not perform as well as the bag-of-words features when evaluated on both general English and biomedical text.

The question of what size window to use prompted the windowing experiment conducted in Chapter 4. The results of this experiment corresponded with [Joshi et al., 2005] and [Gale et al., 1992] findings, although the increase in accuracy when using the larger window size was only by two percentage points indicating that the CUIs closest to the target word are able to classify a majority of the instances.

Syntactic Features

This section discusses the syntactic features shown in Table 7.2. The related work in this table is classified into two sections, those that used the features to disambiguate words in general English text and those that used them to disambiguate words in biomedical text.

Table 7.2: Syntactic WSD Features

General English	Morphology	POS	Head word	Syntactic Relations
[McRoy, 1992]	x			x
[Bruce and Wiebe, 1994]	x	x		
[Ng and Lee, 1996]	x	x		x
[Lee and Ng, 2002]	x	x		x
[Pedersen and Bruce, 1997]	x	x	x	
[Pedersen and Bruce, 1998]	x	x	x	
[Mohammad and Pedersen, 2004]		x	x	
[Yarowksy and Florian, 2003]			x	
[Yarowsky, 1993]				x
[Yarowsky, 1995]				x
[McCarthy, 1997]				x
[Stevenson and Wilks, 2001]				x
Biomedical	Morphology	POS	Head word	Syntactic Relations
[Leroy and Rindflesch, 2004]		x	x	

The syntactic features described in this section include: morphology, part-of-speech (POS), head words and syntactic relations. [McRoy, 1992] defines morphology as the "analysis of each word into its root and affix". Determining the morphology of a word depends on its part-of-speech, if the word is a noun, the morphology feature indicates whether the noun is singular or plural, and if the word is a verb, the morphology feature indicates the tense of the verb.

The head word feature consists of whether or not the target word is the head word in its respective phrase. If one concept of the target word is always used as the head of its phrase while the other not, this becomes a good indicator of what concept is being referred to.

Knowing the POS of a target word narrows down the number of possible concepts of a target word. For example, the word *train* can be either a noun or a verb. If *train*

is being used as a noun some of the possible concepts are:

- train, railroad train
- caravan, train, wagon train
- a series of consequences wrought by an event
- piece of cloth forming the long back section of a gown
- gearing, gear, geartrain, power train

If *train* is being used as a verb, some of the possible concepts are:

- train, develop, prepare, educate
- aim, take, train, take aim, direct
- coach
- exercise in order to prepare for an event or competition
- cause to grow in a certain way by tying and pruning it

Knowing the POS of the target word can reduce the number of possible concepts to choose from. There are a variety of ways head and POS information have been investigated. [Bruce and Wiebe, 1994] use the morphology of the target word and the POS of the words surrounding the target word, [Ng and Lee, 1996] and [Lee and Ng, 2002] use the POS of the target word and the three words to its right and left in their supervised method, and [Pedersen and Bruce, 1997] and [Pedersen and Bruce, 1998] use the POS of the target word and the two words to its right and left in their clustering method. [Leroy and Rindflesch, 2004] use the POS of a target word and whether the target word is a head word. [Mohammad and Pedersen, 2004] compare the following combinations of POS and head word features:

- POS of the specified surrounding words
- the head word of the phrase containing the target word
- the head word of the target word's parent phrase
- the POS of the head word of the phrase containing the target word
- the POS of the head word of the target word's parent phrase

They found that using the POS of the target word plus the POS of the two words to the right and left of the target word obtained the highest disambiguation accuracy.

Syntactic relations are the relationships between the POS of the target word and the POS of its surrounding words. [McRoy, 1992] and [Yarowsky, 1993] use the term *collocation* to refer to syntactic relations. [Yarowsky, 1993] defines a collocation as "two words in some defined relationship" investigating verb-object, subject-verb, and adjective-noun pairs. [McRoy, 1992] defines a collocation as the "relationship among any group of words that tend to co-occur in a predictable configuration". The author goes on to state that collocations are best recognized by their syntactic form. This dissertation refers to these type of collocations as syntactic relations so as not to confuse them with lexical collocations.

The syntactic relations serve as a proxy for selectional restrictions. For example, given the two instances below:

- His class covered 19th century art.
- He covered the boat with a tarp.

The verb-object *cover-art* indicates that art is going to be discussed whereas *cover-boat* indicates the act of spreading something over an object to conceal it. The assumption is that certain concepts of a verb will only take certain objects and therefore the verb can be disambiguated based on its object.

[Ng and Lee, 1996] use the verb-object syntactic relation and define a verb-object syntactic relation to exist 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] use the syntactic relation between the POS of the target word and the surrounding words that are nouns.

[Yarowsky, 1995], [McCarthy, 1997], and [Stevenson and Wilks, 2001] use *selectional preferences* as indicators to determine the appropriate concept of a target word in general English. Selectional preference is a set of restrictions on co-occurring words. [Resnik, 1993] defines this as the entropy between the prior distribution of the possible syntactic relations and the posterior distribution of the relation given the target word.

Semantic Features

This section discusses the semantic features shown in Table 7.3. In this section, the features consists only of those that have been evaluated on general English text. Semantic features are domain dependent and therefore the biomedical features discussed below may be classified as semantic if this dissertation did not distinguish between general English features and biomedical features.

Table 7.3: Semantic WSD Features

General English	Subject Codes	Similarity Measures
[Black, 1988]	x	
[Yarowsky, 1992]	x	
[Stevenson and Wilks, 2001]	x	
[Banerjee and Pedersen, 2003]		x
[Altintas et al., 2005]		x
[Pedersen et al., 2005]		x

Subject codes are a broad categorization of a concept. For example, the term *bat*, when referring to the mammal that flies in the air, has the subject code $Mammal$ whereas the term *bat* when referring to a baseball bat has the subject code $Artifact$. [Black, 1988] uses subject codes from the Longman Dictionary of Contemporary English (LDOCE) as features in their supervised method. [Yarowsky, 1992] and [Stevenson and Wilks, 2001] use the subject codes from Roget's Thesaurus referred to as categorizes.

Similarity and relatedness measures assign a score to how similar or related two concepts are to each other. For example, if the term *bat* is used to refer to the *mammal* then the context surrounding the target word would contain words similar to this concept of bat, such as sonar, night, and insects. These measures have been applied to WSD by obtaining the relatedness or similarity score between the words surrounding the target word and each possible concept of the target word. The scores are combined for each concept and the target word is assigned the concept with the highest score. [Banerjee and Pedersen, 2003], [Altintas et al., 2005], and [Pedersen et al., 2005] evaluate similarity and relatedness measures on the task of WSD.

7.1.2 Biomedical Features

Biomedical features have recently been introduced to disambiguate words specifically in the biomedical domain. These features attempt to capture biomedical knowledge that is not inherent in the text but known within the domain. Table 7.4 shows the biomedical features discussed in this section.

Table 7.4: Biomedical WSD Features

	semantic types	semantic relations	MSH terms	MetaData
[Leroy and Rindflesch, 2004]	x	x		
[Humphrey et al., 2006]	x			
[Fan and Friedman, 2008]	x			
[Stevenson et al., 2008]			x	
[Savova et al., 2008]			x	x
[Alexopoulou et al., 2009]	x			x

A semantic type is a broad subject categorization assigned to a CUI in the UMLS. These are similar to the subject codes from LDOCE or Roget's Thesaurus which is a categorization of general English words. For example, the word *bat* is assigned the subject code *Mammal* in LDOCE and the semantic type *Mammal* in the UMLS.

[Leroy and Rindflesch, 2004] were the first to incorporate biomedical features to disambiguate words in biomedical text. They use the semantic types of the words surrounding the target word as features to their supervised WSD method.

[Humphrey et al., 2006] also incorporate semantic type information. They use the terms associated with the semantic type of a possible concept to create a concept vector in their knowledge-based method. This method determines the appropriate semantic type of target word with the assumption that its possible concepts have a unique semantic type. Using this assumption, [Fan and Friedman, 2008] propose a supervised method to identify the semantic type of a target word. Rather than using human annotated training data, they automatically create training data for a supervised learning algorithm using MetaMap.

A semantic relation is a relationship between semantic types. [Leroy and Rindflesch, 2004] investigate using the semantic relations between the target word and the surrounding words as well as the relations between the surrounding words themselves. They found that using the semantic types of the words surrounding

the target word obtain a higher disambiguation accuracy than using the semantic relations between the words and the target word as well as the semantic relations between the surrounding words themselves.

[Alexopoulou et al., 2009] introduce their *Closest Sense* method which calculates the average of the shortest distance between the semantic type of the possible concept and the semantic types of each of the words surrounding the target word creating a *semantic distance* score for each possible concept. The concept with the lowest semantic distance score is assigned to the concept. This method is similar to the methods that incorporate semantic similarity and relatedness measures.

[Stevenson et al., 2008] were the first to propose using MSH headings as a feature in their supervised method. MSH headings are a set of concepts from the Medical Subject Heading vocabulary which has been incorporated into the UMLS. These headings are manually assigned to biomedical citations in PubMed for indexing purposes. The NLM-WSD dataset consists of abstracts from PubMed where each abstract contains at least one MSH heading.

MetaData consists of terms from different subsections of an article, clinical note or paper. [Alexopoulou et al., 2009] use the following Metadata in their supervised method:

- title
- sentence
- entire abstract
- publication period
- journal title

[Savova et al., 2008] also use Metadata in their supervised method when disambiguating words in clinical text. They incorporate the following data:

- section heading of the clinical note
- medical specialty of the clinical note

In the biomedical domain, researchers use both the general English and biomedical features. Table 7.5 shows the different classification of features that have been used in systems designed to disambiguate words in the biomedical text. [Joshi et al., 2005]

used only general English features in their method whereas [Humphrey et al., 2006] and [Fan and Friedman, 2008] used only biomedical features. The remaining researchers used a mixture of both biomedical and general English features.

Table 7.5: WSD Features

	lexical	syntactic	biomedical
[Joshi et al., 2005]	x		
[Alexopoulou et al., 2009]	x		x
[Savova et al., 2008]	x	x	x
[Liu et al., 2004]		x	x
[Leroy and Rindflesch, 2004]		x	x
[Stevenson et al., 2008]		x	x
[Humphrey et al., 2006]			x
[Fan and Friedman, 2008]			x

7.2 WSD Methods

There are a number of different types of methods that use the features discussed above. This section discusses the methods that have been previously used in general English that have been later evaluated in the biomedical domain or created specifically to disambiguate words in biomedical text. These methods are classified into three categories: supervised methods, clustering methods, and knowledge-based methods. A general description of each of these methods is discussed in Section 2.

An attempt to compare the related work within their respective methods is conducted although not across methods. It is difficult to compare within methods much less across them. There does not exist a centralized evaluation and over the years people have used different subsets and baselines for evaluation. Just within biomedical [Liu et al., 2004], [Leroy and Rindflesch, 2005], [Joshi et al., 2005], [Humphrey et al., 2006], [Savova et al., 2008], and [Fan and Friedman, 2008] use different subsets of the NLM-WSD dataset to evaluate their method.

7.2.1 Supervised WSD Methods

This section discusses previously proposed supervised WSD methods. These methods use manually annotated training data as input into a supervised learning algorithm to

create a supervised learning model which assigns concepts to unannotated test data. A general description of supervised methods is described in Section 2.2.1.

Supervised Learning Algorithms

A number of different supervised learning algorithms have been compared by researchers, some evaluated on general English text while others on biomedical text. Table 7.6 shows the supervised learning algorithms and the researchers who evaluated them.

The following researchers compare supervised learning algorithms, evaluating the accuracy of the algorithms using text from the general English domain. [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 et al., 1993] compare Naive Bayes, Neural Network and Content Vector on the same data set. Both report that the Naive Bayes and Neural Networks return comparable results and each obtains a higher disambiguation accuracy than the other algorithms.

[Yarowksy and Florian, 2003] compare Naive Bayes, Decision Lists, Cosine, Transformation based algorithm and the Bayes Ratio proposed by [Gale et al., 1992]. They report that the Naive Bayes and Bayes Ratio obtain a higher disambiguation accuracy than the other algorithms.

[Ng, 1997] compares Naive Bayes and an improved version of the exemplar-based learning algorithm called LEXAS originally proposed by [Ng and Lee, 1996]. They evaluate their method on a subsection of the British National Corpus (BNC) and a subsection of the Penn Treebank Wall Street Journal (WSJ6) corpus reporting no significant difference in the accuracy of the algorithms. Later, [Lee and Ng, 2002] compare Naive Bayes, Support Vector Machines (SVM), AdaBoost and Decision Trees reporting that SVM obtains the highest disambiguation accuracy.

A similar evaluation of supervised learning algorithms has been conducted in the biomedical domain. [Liu et al., 2004] compare the Naive Bayes, a modified Decision List algorithm and their *mixed* supervised method which is a combination of the Naive Bayes and an exemplar-based algorithms. They report that the Naive Bayes returns a higher accuracy in the general English domain and their *mixed* method returns a higher accuracy in the biomedical domain.

[Joshi et al., 2005] and [Lee and Ng, 2002] compare Naive Bayes, Support Vector

Machines (SVM), AdaBoost, Decision Trees and Decision Lists reporting that SVM returns the highest overall disambiguation accuracy. [Stevenson et al., 2008] compare the Naive Bayes, the SVM and the Vector Space Model (VSM) reporting that VSM returns the highest accuracy with SVM reporting the second highest.

Although, each researcher reported what algorithm obtained the highest disambiguation accuracy they also each noted that no one algorithm performed the best over all the target words in their dataset. The continual use of Naive Bayes and later SVMs to disambiguate words in both the general English and biomedical domains is the reason the K-CUI experiments use these algorithms for evaluation in Chapter 4.

Supervised WSD Results

This subsection discusses the results of the different features that have been previously used in supervised WSD methods to disambiguate words in the general English and biomedical domain. Table 7.7 shows the researcher, features used in their proposed supervised method, the overall disambiguation accuracy of the system, and the dataset used for evaluation for methods evaluated in the general English domain. Table 7.8 shows similar results for those methods evaluated in the biomedical domain.

[Leacock et al., 1993], [Mooney, 1996] and [Pedersen, 2000] evaluate the bag-of-words feature set using the *line* data reporting an accuracy of 76%, 72%, and 88% respectively. [Mohammad and Pedersen, 2004] report a higher accuracy using unigrams and bigrams on the same dataset.

[Ng and Lee, 1996] evaluate bag-of-words and collocations on the *interest* data set finding that bag-of-words performed lower than collocations. The collocation results are higher than the results reported by [Mohammad and Pedersen, 2004] using unigram but not bigrams.

[Lee and Ng, 2002] evaluate unigrams and collocations on the SENSEVAL-1 and SENSEVAL-2 datasets finding unigrams obtain a lower disambiguation accuracy than collocations. The collocation results are also higher than the SENSEVAL-1 and SENSEVAL-2 unigram and bigram results reported by [Mohammad and Pedersen, 2004] who show that there is no difference in bigram and unigram results. [Joshi et al., 2005] evaluate unigrams and bigrams on biomedical text finding that the length of the sentences were to small to identify significant bigrams in most cases which may explain why their is

Table 7.6: Supervised WSD Methods

	Mooney 1996	Leacock and Chodorow 1998	Yarowsky and Florian 2003	Liu, et. al. 2004	Ng and Lee 1996	Ng 1997	Lee and Ng 2002	Josh, et. al. 2005	Leroy and Rindflesch 2005	Lee, et. al. 2005	Stevenson, et. al. 2008	Fan and Friedman 2008	Savova, et. al. 2008
Naive Bayes	x	x	x	x		x		x	x	x	x	x	
SVM							x	x		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/NB *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												
Vector Space Model											x		

Table 7.7: Supervised WSD Results Evaluated on General English Text

		SENSEVAL-1	SENSEVAL-2	*line*	*hard*	*serve*	*interest*
[Leacock et al., 1993]	bag-of-words			76			
[Mooney, 1996]	bag-of-words			72			
[Pedersen, 2000]	bag-of-words			88			
[Ng and Lee, 1996]	bag-of-words						62
	collocations						80
	POS + morphology						77
	verb-object relation						77
[Mohammad and Pedersen, 2004]	unigrams	67	55	75	83	73	76
	bigrams	67	55	73	89	72	79
	POS	68	55	60	85	76	80
	Head	64	52	55	88	47	69
	Head of Parent	60	50	60	85	57	68
	POS of Phrase	59	53	54	82	41	55
	Parent Phrase POS	58	53	54	82	42	55
[Lee and Ng, 2002]	unigrams	70	58				
	collocations	74	61				
	POS	70	55				
	syntactic relation	70	55				

very little difference in [Mohammad and Pedersen, 2004] unigram and bigram results.

[Lee and Ng, 2002] evaluate using the POS of the surrounding words and noun syntactic relations on the SENSEVAL-1 and SENSEVAL-2 datasets finding that each feature set obtains a similar disambiguation accuracy. In their early work, [Ng and Lee, 1996] evaluate a combination of POS and morphology features and verb-object on the *interest* dataset showing that the combination of POS and morphology obtains a higher disambiguation accuracy than verb-object relations.

[Mohammad and Pedersen, 2004] evaluate their method using various POS information. The authors report that the POS of the surrounding words obtains 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. They also report 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.

Table 7.8: Supervised WSD Results Evaluated on Biomedical Text

		NLM-WSD						Clinical
		Leroy	Liu	Joshi	Savova	Fan	All	Dataset
[Leroy and Rindflesch, 2004]	Head	58						
	POS	58						
	ST (phrase)	61						
	ST (sentence)	66						
	NC relation	55						
	C relation	56						
	NC Sense Act.	60						
[Joshi et al., 2005]	bigrams	77	85	83				
[Liu et al., 2004]	combination		86					
[Stevenson et al., 2008]	Linguistic+MSH	79	85	83	86	88	88	
[Savova et al., 2008]	Linguistic+MSH	79	85	81	86			
	Linguistic+MetaData							83.8
[Fan and Friedman, 2008]	ST					71		

[Leroy and Rindflesch, 2004] evaluate their system on a subset of the NLM-WSD data set (Leroy subset) using combinations of the following features:

- head word information

- POS of the target word
- semantic types of the words surrounding the target word
- semantic relations between the target word and the surrounding words
- semantic relations between the surrounding words themselves

The authors report a small improvement in accuracy over the baseline using whether the target word was a head word (Head) as a feature but no substantial increase or decrease in performance when using the target word's POS. When adding the semantic types of the words that occur in the phrase as the target word decreases the results but including the semantic type of the words in the same sentence as the target word increases them. They authors also add the semantic relations between the semantic type of the target word and each of the words but showed that it did not improve the accuracy of the results.

[Liu et al., 2004] evaluate their system on a different subset of the NLM-WSD dataset (Liu subset). They utilize combinations of the following features:

- bag-of-words
- unigrams
- orientation
- distance
- collocations
- unigrams.

They compare the Naive Bayes, a modified Decision List and a combination Naive Bayes/exemplar-based algorithm and report the best per word accuracy over all feature sets and algorithms for each target word in the subset.

[Joshi et al., 2005] evaluate their method on the Leroy and Liu subsets, and their union referred to as the Joshi subset. They report that using the unigrams obtains a higher disambiguation accuracy than the best results reported by [Leroy and Rindflesch, 2004] and are comparable to the results reported by [Liu et al., 2004].

[Stevenson et al., 2008] evaluate their method on the entire NLM-WSD dataset using the following features:

- collocations

- syntactic dependencies
- bag-of-words combined
- MSH headings

The authors evaluate their system on the entire NLM-WSD dataset as well as the subsets obtaining an higher disambiguation accuracy than [Leroy and Rindflesch, 2004], [Joshi et al., 2005] and [Liu et al., 2004] on their respective subsets.

[Savova et al., 2008] evaluate their method a subset of the NLM-WSD dataset referred to as the Savova subset and a dataset of clinical notes referred to as Clinical. They evaluate combinations of the following features:

- bag-of-words
- stemmed words
- POS,
- words within a window size of five, 10 and 50 of the target word
- orientation of the words with respect of the target word
- how far from the target word the word exists
- MSH headings
- named entities
- Metadata

They report the accuracy of the best combination of features for each target word individually.

[Fan and Friedman, 2008] evaluate their method using a subset of the NLM-WSD dataset referred to as the Fan subset. This method actually identifies the semantic type of the target word rather than its concept. The method then uses this information to determine which concept to assign to the target word with the assumption that each possible concept of a target word has a unique semantic type. The semantic types are a more coarse grained categorization than CUIs which makes them easier to assign. The advantage of this method is that it does not use manually annotated training data but MetaMap mapped text as training data instead. The overall disambiguation accuracy is 71% which is lower than the other supervised methods but is a more practical for real world applications because it does not require manually annotated training data.

7.2.2 Clustering WSD Methods

This section discusses previously proposed clustering WSD methods. These methods originates from work conducted in the area of Information Retrieval (IR). An example of an early IR method which uses clustering is by [Salton et al., 1975] who introduce a method to automatically index documents for retrieval. This method treats documents as vectors and clusters them in an n-dimensional space. A new document vector is compared to each of the clusters centroid. The cluster whose centroid is closet to the new document vector is returned. [Schütze, 1992] proposed this type of method for word sense discrimination which seeks to cluster instances of a given target word such that instances that use the same concept of the target word are in the same cluster. In this method, [Schütze, 1992] introduces using second-order vectors rather than first order vectors in order to alleviate the sparseness in the first-order vectors.

There are a number of different features that have been used to create these vectors. Table 7.9 shows the different features that have been used to create the training vectors for the clustering method. [Pedersen and Bruce, 1997] and [Pedersen and Bruce, 1998] use three different feature sets:

- 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,
- morphology, unrestricted collocations with the two words to the left of the target word and right of the target word, and
- 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.

The authors note that the feature sets used are better at distinguishing between noun concepts than verb and adjective concepts. [Purandare and Pedersen, 2004] evaluate the second-order co-occurrence proposed by [Schütze, 1998] and compare them with first-order co-occurrences, first-order bigrams and second-order bigrams.

Table 7.10 shows the results of reported for each of the different features discussed above. [Pedersen and Bruce, 1997] and [Pedersen and Bruce, 1998] evaluate their method using the *line* data and the [Bruce and Wiebe, 1994] corpus ($Mix1$). Their

Table 7.9: Clustering WSD Methods

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

results both show that the feature set containing Morphology(M), POS of the surrounding words and the content collocations obtained the highest accuracy regardless of the clustering algorithm.

Table 7.10: Clustering Results

	features	*Mix*1	*line*	*hard*	*serve*	SENSEVAL-2LS
	M + POS + First-order Co-occurrence	65.5				
Pedersen and	M + Unrestricted Collocations	65.3				
Bruce, 1997	M + POS + Content Collocations	66.2				
	M + POS + First-order Co-occurrence		64.6			
Pedersen and	M + Unrestricted Collocations		65.7			
Bruce, 1998	M + POS + Content Collocations		65.9			
	First-order Collocations		62	41	37	44
Purandare and	First-order Bigrams		68	87	46	44
Pedersen, 2004	Second-order Collocations		55	73	34	43
	Second-order Bigrams		38	63	31	47

[Purandare and Pedersen, 2004] evaluate their method using the *line*, *hard*, *serve* and SENSEVAL-2LS data sets. They show that first-order bigram vectors obtain a higher disambiguation accuracy when evaluated on the *line*, *hard*, and *serve* data sets while second-order bigram vectors obtain a higher disambiguation accuracy when evaluated on the SENSEVAL-2LS data set. The authors note that SENSEVAL-2LS is a smaller data set than the others which may be reflected in the results.

7.2.3 Knowledge-based WSD Methods

This section discusses previously proposed knowledge-based WSD methods. These methods use information extracted from a knowledge source to determine the appropriate concept of a target word. There methods discussed in this section are broken into two categories: those that use the location of the vectors in some n-dimensional space (knowledge-based vector methods) and ii) those that use semantic similarity measures (knowledge-based similarity methods). A general description of knowledge-based methods is described in Section 2.2.3.

Knowledge-based Vector Methods

The methods discussed in this section use the location of the test and concept vectors in some n-dimensional space to determine the correct concept of the target word. These methods are very similarly to A-CUI. In A-CUI, a test vector and a concept vector for each possible concept of the target word are created using first or second-order vectors. A measure is used to determine the distance between the test vector and each of the concept vectors. The concept whose vector is the closest to the test vector is assigned to the target word. The context used to create the test vector comes from the words surrounding the target word while the context used to create the concept vectors come from information about the concepts CUI in the UMLS or the 2005 Medline baseline. The methods discussed in this section have all been evaluated on different datasets and are therefore incomparable.

[Humphrey et al., 2006] introduce a similar method that uses semantic type information. Their method determines the appropriate semantic type of a target word with the assumption that each of the possible concepts of the target word have a unique semantic type. For example, consider the target word *culture* which has two possible concepts, Anthropological Culture [C0010453] and Laboratory Culture [C0430400], each with a different semantic type. The semantic type for Anthropological Culture [C0010453] is *Idea or Concept* while the semantic type for Laboratory Culture [C0430400] is *Laboratory Procedure*.

In this method, a first-order *concept* vector is created for all of the concepts of the target word. The feature set contains all of the one word terms in the Metathesaurus.

The elements in the vector indicate whether or not the CUI of the term is assigned that semantic type. A test vector is then created whose elements indicate whether or not a feature is one of the semantic types of the words surrounding the target word. The angle is then calculated between the test vector and each of the concept vectors using the Cosine Measure. The concept of the concept vector closest to the test vector is assigned to the target word. The disadvantage of this method is that if two possible concepts have the same semantic type the system is not be able to disambiguate between them. 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 Rindflesch, 2004] who obtained a 65.6% accuracy evaluating their supervised WSD system.

[Mohammad and Hirst, 2006] introduce a similar method that uses information from Maquarie's Machine Readable Thesaurus. In this method, a concept vector is created for each possible category in Maquarie's. The feature set contains all of the words in the thesaurus and the elements indicate whether or not the feature occurs with the category in the British National Corpus. A test vector is created whose elements indicate whether or not the feature exists in the words surrounding the target word. The value of the non-zero elements in the vectors is the measure of association between the category and the feature. The authors investigate four measures of association:

- Dice Coefficient
- Cosine Measure
- Pointwise Mutual Information
- Odds Ratio
- Yule's Coefficient
- Phi Coefficient

The distance between the test vector and each of the concept vectors is calculated using a *Dominance* metric. The *Dominance* of a category is then calculated based on the association scores from the vectors. 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*:

- the normalized sum of the association scores between the surrounding words and the category

- the maximum association of the surrounding words and the category divided by the number of surrounding words
- 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
- 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 report an overall accuracy results for the entire dataset. They instead reported the accuracy of word groupings based on the number of possible concepts, for example, words that contain only one concept are grouped together, those that contain two are grouped together. The authors showed that *Dominance* calculations one and three outperform calculations two and three. They also showed that the Odds Ratio, Point Wise Mutual Information and Yule's Coefficients obtained higher disambiguation accuracies than the other measures.

[Patwardhan, 2003] introduces a semantic relatedness measure called UMLS::SIMILARITY::VECTOR. This measure determines the semantic relatedness between two concepts in WordNet. The measure is described here because of its closeness to A-CUI and the WSD methods proposed by [Humphrey et al., 2006] and [Mohammad and Hirst, 2006]. UMLS::SIMILARITY::VECTOR calculates the semantic relatedness between two concepts in WordNet by creating a second-order concept vector for each concept using its definition (or gloss as it is called in WordNet) and the definition of its related concepts as context. The cosine measure is then used to calculate the angle between the two vectors and this score is used to quantify their similarity. The assumption behind this measure is that the definitions provide contextual information about a concept and similar concepts will have similar contextual information.

The similarity between A-CUI and the methods described in this section is that a concept vector is created using information extracted about the concept from a knowledge-source. The source of knowledge and how the vector is created is what differentiates these methods.

Knowledge-based similarity Methods

Semantic similarity and relatedness measures have been applied to the task of word sense disambiguation. This is obtained by calculating the relatedness or similarity score between the words surrounding the target word and each possible concept of the target word. The scores are then either summed or averaged for each concept and the target word is assigned the concept with the highest score. The lexical database used by the related work in this section is WordNet[1] unless specified.

Similarity measures are categorized as: path-based measures and information content (IC) measures. Path-based measure rely solely on the location of the concepts in a taxonomy such as those proposed by:

- [Rada et al., 1989]
- [Leacock and Chodorow, 1998]
- [Altintas et al., 2005]
- [Wu and Palmer, 1994]
- [Agirre and Rigau, 1996]
- [Hirst and St-Onge, 1998]

IC measures rely on the probability of a concept occurring such as those proposed by:

- [Resnik, 1995]
- [Jiang and Conrath, 1997]
- [Lin, 1997]

A distinction is made between similarity and relatedness measures because concepts that are not similar can still be related. For example, *silicon* and *computers* are related but not similar. Therefore, the similarity score between them would be low while their relatedness score would be high. Examples of relatedness measure are those proposed by [Lesk, 1986], [Cowie et al., 1992], and [Patwardhan, 2003]. A more detailed analysis of similarity and relatedness measures can be found in Appendix I.

Semantic similarity measures have been evaluated on the task of WSD. [Agirre and Rigau, 1996] evaluate their measure on a four texts from the SemCor dataset: br-a01 (Press:Reportage), br-b20 (Press:Editorial), br-j09 (Learned:Science)

[1] WordNet is described in Section 2.3.1

Table 7.11: Analysis of Semantic Similarity Measures Applied to WSD

		Agirre and Ragu 1996	Banerjee and Pedersen 2003	Altintas, et. al. 2005	Pedersen et. al. 2005
Path-based Similarity Measures	[Leacock and Chodorow, 1998]		32	32/97	23
	[Wu and Palmer, 1994]			33/97	30
	[Hirst and St-Onge, 1998]				20
	[Agirre and Rigau, 1996]	70			
	[Altintas et al., 2005]			35/97	
	[Sussna, 1993]	65			
IC Similarity Measures	[Resnik, 1995]		30	30/83	29
	[Jiang and Conrath, 1997]		38	31/72	40
	[Lin, 1997]		31	38/58	29
Relatedness Measures	[Lesk, 1986]				28
	[Banerjee and Pedersen, 2003]		39		41
	[Patwardhan, 2003]				29
	Metric	Accuracy		PR	Fmeasure
	Data Set	SemCor subset	SENSEVAL-2LS nouns dataset		

and br-r05 (Humor). The authors compare their results to their implementation to [Sussna, 1993] obtaining a higher accuracy. The authors also compare their method to proposed by [Yarowsky, 1992] obtaining a higher overall disambiguation accuracy.

[Banerjee and Pedersen, 2003] evaluate their measure on the noun data from the SENSEVAL-2LS dataset reporting their results using accuracy. The authors compare their measure with those proposed by [Leacock and Chodorow, 1998], [Resnik, 1995], [Jiang and Conrath, 1997], [Lin, 1997] and [Hirst and St-Onge, 1998] using the SenseRelate[2] software package which uses the WORDNET::SIMILARITY[3] module to determine the similarity between two concepts. The probability information for the information content measures comes from a combination of SemCor, the Brown Corpus, the Penn Treebank and the British National Corpus. The results show that [Jiang and Conrath, 1997] obtain a higher accuracy (38%) than the other proposed measure but their measure obtains the highest accuracy (39%).

[Altintas et al., 2005] also evaluate their measure on the noun data from the

[2] http://sourceforge.net/projects/senserelate/
[3] http://search.cpan.org/dist/WordNet-Similarity/

SENSEVAL-2LS dataset reporting their results using precision and recall (P/R). The authors compare their measure with those proposed by [Leacock and Chodorow, 1998], [Wu and Palmer, 1994], [Resnik, 1995], [Jiang and Conrath, 1997] and [Lin, 1997] using the WORDNET::SIMILARITY module. The probability information for the information content measures comes from the SemCor dataset (the default setting of WORDNET::SIMILARITY). The authors report scores similar to that of [Banerjee and Pedersen, 2003]. They report that the measure proposed by Lin obtained the highest precision (38%) but the lowest recall whereas their measure obtained the second highest precision (35%) and the highest recall (97%) tying with the measures proposed by [Leacock and Chodorow, 1998], and [Wu and Palmer, 1994].

[Pedersen et al., 2005] evaluate the measures proposed by [Banerjee and Pedersen, 2003] and [Patwardhan, 2003] on the noun, verb and adjective datasets from the SENSEVAL-2LS dataset reporting their results using the F-measure. The authors compare the two measures with those proposed by [Leacock and Chodorow, 1998], [Wu and Palmer, 1994], [Hirst and St-Onge, 1998], [Resnik, 1995], [Jiang and Conrath, 1997], [Lin, 1997] and [Lesk, 1986] using the SenseRelate. They explore using a varying number of words surrounding the target word as context (referred to as window size) concluding in general the more context that is used the better the results. The authors also evaluate obtaining the probability information for the information content measures from SemCor and the British National Corpus. Table 7.11 shows the F-measure for the noun data from the SENSEVAL-2LS using the probability information derived from SemCor using a window size of 11. The authors report scores consistent to Banerjee and Pedersen, and Altintas et. al. The results show that [Jiang and Conrath, 1997] (40%) obtain a higher F-measure than the other proposed measure but the measure proposed by [Banerjee and Pedersen, 2003] obtains the highest F-measure (41%).

Chapter 8

Future Work

The overall results of the A-CUI and K-CUI experiments show that CUIs provide a unique source of information about a possible concept which can be used in both supervised and knowledge-based word sense disambiguation systems. K-CUI uses the CUIs of the words surrounding the target word as features into a supervised learning algorithm. A-CUI uses the CUI's definition and words surrounding the CUI as contextual information describing that CUI to create a concept vector for a knowledge-based algorithm.

K-CUI employs both an MMI cutoff and a semantic similarity cutoff to determine which CUIs to include in the feature set. The results showed that both cutoffs obtained a comparable disambiguation accuracy to that of not using a cutoff while significantly reducing the number of features in the feature set. In the future, plans exist to evaluate this dataset on other types of biomedical text such as clinical notes.

Using a similarity cutoff showed that the number of features included in the feature set reduced by over half. The similarity between the two concept was calculated by the UMLS::Similarity package using measures that rely on the path information between the two concepts. This information was obtained using SNOMED-CT. The disadvantage of this is that not all of the possible concepts in the NLM-WSD dataset exist in SNOMED-CT. In the future, plans exist to expand the UMLS::Similarity so that the path information can be obtained for all of the possible concepts in the dataset. Plans also exist to investigate using K-CUI with these two cutoffs on biomedical text other than the NLM-WSD dataset such as clinical text.

The success of using semantic similarity in K-CUI begs the question of how this type of information can be incorporated into A-CUI. In the future, plans exist to create the test and concept vectors using CUIs as features rather than highly frequent words. The elements in the first-order concept vectors would be either a one or a zero indicating whether or not the feature, the CUI, and the concept have a semantic similarity score higher than a specified threshold. The second-order concept vector is created by first creating a first-order vector for the CUIs that have a semantic similarity score higher than a specified threshold with the concept. The features in the first-order vectors would be the CUIs in the UMLS and the elements would be the either a one or a zero indicating if the CUI occurred with the feature in the 2005 Medline baseline. One potential disadvantage of creating a second-order concept vector like this is that amount of computational time it will take to create it.

The elements in the vectors are currently binary. In the future, plans exist to explore using the semantic similarity score instead. For example, the elements in the first-order test vector would be the semantic similarity score between the concept and the feature rather than a binary indicator of their presence or absence in the 2005 Medline baseline.

There are also plans to implement the knowledge-based similarity method such as the SenseRelate[1] system proposed by [Pedersen et al., 2005] using the UMLS rather than WordNet as the knowledge-source. In this method, for each possible concept of a target word the similarity score calculated between it and the words surrounding it. These scores are summed creating a single score for the concept. The concept with the highest similarity score is assigned to the target word.

The future work discussed in this chapter focuses on incorporating semantic similarity into A-CUI. As of version 0.13, the UMLS::Similarity package only contains path-based measures. There are other similarity measures such as the information content (IC) measures proposed by [Resnik, 1995], [Jiang and Conrath, 1997] and [Lin, 1997] which incorporate the probability of a concept occurring. [Pedersen et al., 2007] showed that the information content measures obtain a higher correlation to humans than the path based measures. In the future, the plan is to investigate if using the IC measures would provide better results than the path measures.

Also, as of version 0.13, that package only contains semantic similarity measures,

[1] http://sourceforge.net/projects/senserelate/

plan exist to expand the package to include relatedness measures. Semantic relatedness is a more general form of semantic similarity. For example, *foot* and *pedal edema* are not similar but are related, where as *foot* and *hand* are both similar and related. In the future, the plan is to compare the results using semantic similarity measures versus semantic relatedness measures to determine if one type is preferable over the other for the purpose of word sense disambiguation.

Chapter 9

Conclusions

This chapter discusses the specific contributions of K-CUI and A-CUI and then provides a general summary of their overall contributions. The specific K-CUI contributions of this dissertation are as follows.

Humans only require a small window size around a target word to determine its appropriate concept. An experiment conducted by [Choueka and Lusignan, 1985] found that only two or three words were needed indicating that only the words closest to the target word are required for disambiguation. [Gale et al., 1992] found this was not the case for supervised WSD methods showing that larger window sizes returned better results when disambiguating words in general English using general English features. The K-CUI results showed a similar finding, using CUIs that occur anywhere in the instance obtains the highest disambiguation accuracy, although using CUIs in the same phrase as the target word is able to disambiguate the target word in a majority of the instances.

This shows in biomedical text indicative features of a concept are highly localized as they are in general English.

The results using a frequency cutoff indicate that CUIs that occur only a few times in the training data are playing a significant role in the disambiguation process. The average number of features in the feature set when not using a cutoff is 2455.48, this decreases to 1426.98 when using a cutoff of two therefore there exists approximately 1028.71 features that only occur once in the training data. The average number of

features that occur in the training data and in the test instance is 50.77 when not using a cutoff, and 44.92 when using a cutoff of two therefore there exists approximately 5.85 features that only occur once in the training data and in the test data. The overall results decrease by only one percentage point using a frequency cutoff of two but analysis of the individual results shows this decrease can be much greater, for example, the accuracy for the target word *lead* decreases from 90% to 83% when the low frequency features are removed.

This indicates that CUIs that only occur a few times in the training data play a significant role in the disambiguation process.

The results using the MMI cutoff show that it is able to significantly reduce the amount of noise in the feature set. When using an MMI cutoff of 10, the feature set contains almost 60% fewer features going from on average 3752.646 features to 1489.92. With the reduced feature set, there are approximately 70% less features seen in a test instance going from on average 63.49 non-zero elements to 18.58. The results show that the overall accuracy only dropped by one percentage point indicating that the MMI cutoff can be used to remove CUIs that are not needed in the disambiguation process while still maintaining a comparable accuracy. The MMI score was created to facilitate an indexing system which recommends MSH headers (CUIs from the MSH vocabulary) to medical text indexers. An MMI score quantifies how relevant a CUI is in describing a Medline abstract therefore a high MMI score indicates that the CUI is useful in describing the overall topic of the abstract.

The success of using an MMI cutoff shows that in biomedical text word senses are correlated with the topical information describing an instance, as they are in general English.

The results using the semantic similarity cutoff also show that it is able to significantly reduce the amount of noise in the feature set. The feature set when using the semantic similarity measure proposed by [Wu and Palmer, 1994] with a cutoff 0.1 contains almost 70% less features going from 3752.64 to 1026.38. With the reduced

feature set, there are approximately 60% less features seen in a test instance going from on average 63.49 non-zero elements in the test vector to 25.62. The system is able to disambiguate the test instances using less features while maintaining a comparable accuracy. A semantic similarity quantifies how "alike" two concepts are by determining their closeness in a hierarchy. [Miller and Charles, 1991] show that the similarity between words can be determined based on the similarity between their contexts indicating that the context surrounding a target word will have a higher similarity score with its correct concept than the other possible concepts. For example, the concept Influenza [C0021400] has a semantic similarity score of 0.1000 with Cold Temperature [C0009264] and 0.2500 with Common Cold [C0009443]. The concept Influenza occurs in only two instances in the NLM-WSD dataset but is still an uniquely representative feature because it is not likely to occur in instances in which *cold* is referring to a concept other than the Common Cold.

This indicates that in biomedical text features that have a high semantic similarity with at least one of the possible concepts of a target word are better able to uniquely represent the context in which the concept is used.

The specific A-CUI contributions of this dissertation are as follows.
Using the definitions of a concept to create a second-order concept vector obtains the highest overall disambiguation accuracy in the A-CUI experiments. A definition is a written explanation of a concept intended for human understanding. It is a precise and tightly focused explanation containing primarily content words that are relevant to the understanding of the concept. For example, the definition of the concept Anthropological Culture [C0010453] is: "A collective expression for all behaviour patterns acquired and socially transmitted through symbols." Second-order vectors are created by first creating a first-order vector for each content words in the definition. The features in the first-order vector come from the training data and the elements are numeric indicators of whether or not the content word occurs with the feature in the training data. These first-order vectors are a contextual representation of the words in the definition which are then averaged together to create a second-order concept vector. This vector provides a contextual representation of the content words in the

definition.

The success of using definitions to create a second-order vector to represent the context of a concept shows that the context used with the words in a concept's definition can be used to represent the context of the concept itself.

The parent definitions provide a broader definition of a concept. The results show that using the parent definition in conjunction with the concept's CUI definition increases the overall disambiguation accuracy. The parent definitions provide additional words describing the concept in general terms. For example, the parent of the concept Aspirin [C0004057] is Antirheumatic Agents [C0003191] which has the definition: Agent that relieves or prevents rheumatic disease, especially rheumatoid arthritis. The definition contains a broader general description of the concept Aspirin [C0004057]. Second-order vectors provide an aggregated contextual representation of the words in the concept's definition as well as the word's in the parent definition. The additional information provided by the parent definition increases the overall disambiguation accuracy indicating that the context used with the words used to describe a concept in general terms can be used in place of the context of the concept.

The success of including the parent definition along with the concept's definition to create a second-order vector to represent the context of a possible concept shows that the context of the words used to describe the general idea of a concept is similar to the context used with the concept itself.

The child definition provides a narrower set of contextual information. For example, the concept Drugs [C1254351] has the child Hormone Antagonists [C0019927] which has the following definition: "Chemical substances which inhibit the function of the endocrine glands, the biosynthesis of their secreted hormones, or the action of hormones upon their specific sites". The accumulation of child definitions provide a general description of the concept. The second-order vectors provide an aggregated contextual representation of the words in the concept's definition as well as the words in the child definitions. The contextual representation of the words in these definitions is used in

replace of the contextual representation of the concept itself. The information in the second-order vector provides the context in which the words in the child definitions are used.

Similarly, this shows that the context used with the words in the child definition is similar to the context used with the concept itself providing additional contextual information about the concept.

The sibling concepts are those concepts that have at least one parent in common. The results show that using this information decreases the overall disambiguation accuracy. The sibling concepts can be very broad for concepts that are higher up the tree, for example, the concept Drugs [C1254351] which has a height of three in the National Cancer Institute Thesaurus has the following sibling concepts:

- Air [C0001861]
- Dust [C0013330]
- Liniment [C0023742]
- Oil [C0028908]
- Ointment [C0028912]
- Soap [C0037392]
- Solution [C0037633]
- Water [C0043047]
- Nail Polish [C0304644]

These concept have seemingly little to do with Drugs [C1254351]. The definition of the concept Air [C0001861] is: A mixture of gases making up the earth‘s atmosphere, consisting mainly of nitrogen, oxygen, argon, and carbon dioxide. The context used with the words in this definition is different than the context that is used with the concept Drugs [C1254351] providing very little useful and potentially harmful

contextual information for the purposes of disambiguation.

This shows that the context used with the words in the sibling definitions is different than the context used with the concept itself.

The UMLS contains very few source synonymous definitions, only one of the possible concepts in the NLM-WSD data has a synonymous relation. The addition of this information decreased the overall accuracy of the results for that target word but with only one extra definition it is difficult to make any clear statement on its behaviour.

This shows that there is not enough source synonym information in the UMLS to be used in providing additional contextual information about a concept.

Using the associated terms of a possible concept as its contextual description did not obtain as high of an overall disambiguation accuracy than using preferred term of the concept. The preferred term for each of the possible concepts of a target word is unique to that concept, whereas an associated term may describe two different possible concepts.

This shows that the associated terms themselves do not provide enough contextual information to disambiguate between them.

The results using the words surrounding the CUI in MetaMap mapped text obtains the second highest overall disambiguation accuracy. Although, using the definitions results in a higher overall disambiguation accuracy there are only a limited number of them that actually exist in the UMLS. Where as, the contextual information, for a majority of the possible concepts in the NLM-WSD dataset can be found in the MetaMap mapped text. MetaMap does not perform WSD but maps concepts to terms. The terms provide enough information for MetaMap to distinguish between the possible concepts but it runs into trouble when the term itself is ambiguous. For example, MetaMap will map the correct concept to the terms “laboratory culture” and “anthropological culture” but can not distinguish between the concepts when the term

just consists of the word "culture" itself. The information extracted about the possible concepts of a target word is unique enough that the words extracted are not always identical.

This indicates that the context, provided by MetaMap, of the possible concepts of a target word is distinct enough to distinguish between them.

The results using the words surrounding the associated terms of a concept in the text did not obtain as high of an overall disambiguation accuracy as using the CUIs. There contains overlap between associated terms of the possible concepts of a target word therefore the context surrounding the terms for each of the possible concepts is not unique.

There contains too much overlap between the associated terms of the possible concepts of a target word to provide a distinct contextual representation for each of the possible concepts.

There are two types of contextual information used to represent a possible concept: definitions and the words surrounding a concept in MetaMap mapped text. The definitions contain human descriptions describing a concept. The MetaMap mapped text contains a contextual description which consists of the words that are commonly used with the concept. The results show that first-order vectors obtain a higher disambiguation accuracy when using the definitions and second-order vectors obtain a higher disambiguation accuracy when using the context provided by MetaMap mapped text.

The definitions do not represent the actual context of how a concept is used in text but a description of the concept. The first order vector of a definition consists of a representation of the definition itself rather than a contextual representation of the concept. Second-order vectors are an aggregation of the contextual representation of the words in the definition.

This shows that second-order vectors should be used when the the information about a concept consists of words that are used to describe the concept rather than the words

that are used with the concept in a text.

The words surrounding the concept in MetaMap mapped text represent the actual context in which a concept is used therefore the first-order vectors are an actual representation of the context. Second-order vectors represent an aggregation of the contextual representation of the words that are used in the same context as the concept. This generalizes the context of the concept too much decreasing the overall disambiguation accuracy.

This shows that first-order vectors should be used when the information about a concept consists of the words surrounding the concept in a text.

The K-CUI and A-CUI methods have been implemented as open source packages[1] that can be used to disambiguate words in any type of biomedical text using information from the UMLS and MetaMap mapped text. K-CUI is a supervised WSD system that uses CUIs assigned by MetaMap to the words surrounding the target word as features into a supervised learning algorithm. The novelty of K-CUI is using MetaMap to map terms to CUIs in the UMLS to be used as features in a supervised system. A-CUI is a knowledge-based WSD system that uses contextual information about a concept from the UMLS and MetaMap mapped text. The novelty of A-CUI is the creation and development of a knowledge-based vector method to determine the appropriate concept of a target using the information from the UMLS and MetaMap mapped text to obtain a contextual description of a concept to use in creating a concept vector.

K-CUI obtains a higher accuracy than A-CUI because it is a supervised system which learns from manually annotated training data. However, this means that it is not practical for real world applications, such as information retrieval systems, because it requires manually annotated training data for each word that needs to be disambiguated.

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

References

[Agirre and Martinez, 2004] Agirre, E. and Martinez, D. (2004). The basque country university system: English and basque tasks. In *Proceedings of the 3rd ACL Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at the Annual Meeting of the Association of Computational Linguistics*, pages 44–48, Barcelona, Spain.

[Agirre and Rigau, 1996] Agirre, E. and Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of the 16th Annual Meeting of the Association for Computational Linguistics*, pages 16–22, Santa Cruz, CA.

[Alexopoulou et al., 2009] Alexopoulou, D., Andreopoulos, B., Dietze, H., Doms, A., Gandon, F., Hakenberg, J., Khelif, K., Schroeder, M., and Wachter, T. (2009). Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics*, 10(1):28.

[Altintas et al., 2005] Altintas, E., Karsligil, E., and Coskun, V. (2005). A new semantic similarity measure evaluated in word sense disambiguation. In *Proceedings of the 15th Nordic Conference of Computational Linguistics*, pages 8–11, Joensuu, Finland.

[Aronson, 1997] Aronson, A. (1997). The MMI ranking function. National Library of Medicine Technical Report. http://skr.nlm.nih.gov/papers/references/ranking.pdf.

[Aronson, 2001] Aronson, A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 17–21, Washington, DC.

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

[Black, 1988] Black, E. (1988). An experiment in computational discrimination of english word senses. *IBM Journal of Research and Development*, 32(2):185–194.

[Bruce and Wiebe, 1994] Bruce, R. and Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, pages 139–146, Las Cruces, NM.

[Budanitsky and Hirst, 2001] Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources at the Annual Meeting for the North American Chapter of the Association of Computational Linguistics*, pages 29–34, Pittsburg, Pennsylvania.

[Budanitsky and Hirst, 2006] Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

[Caviedes and Cimino, 2004] Caviedes, J. and Cimino, J. (2004). Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85.

[Choueka and Lusignan, 1985] Choueka, Y. and Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, 19(3):147–157.

[Cowie et al., 1992] Cowie, J., Guthrie, J., and Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 359–365, Nantes, France.

[Fan and Friedman, 2008] Fan, J. and Friedman, C. (2008). Word sense disambiguation via semantic type classification. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 177–181, Washington, DC.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet – An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts and London, England.

[Fix and Hodges, 1951] Fix, E. and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination. usaf school of aviation medicine, randolph field, tx. Technical report, Project 21-49-004, Report 4, Contract AF41 (128)-3, 1951.

[Gale et al., 1992] Gale, W., Church, K., and Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439.

[Graziano et al., 1997] Graziano, G., Catanzano, F., Riccio, A., and Barone, G. (1997). A reassessment of the molecular origin of cold denaturation. *Journal of Biochemistry*, 122(2):395.

[Hirst, 1987] Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, England.

[Hirst and St-Onge, 1998] Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An Electronic Lexical Database*, pages 305–332.

[Humphrey et al., 2006] Humphrey, S., Rogers, W., Kilicoglu, H., Demner-Fushman, D., and Rindflesch, T. (2006). Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technolology*, 57(1):96–113.

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

[Joshi et al., 2005] Joshi, M., Pedersen, T., and Maclin, R. (2005). A comparative study of support vectors machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of 2nd Indian International Conference on Artificial Intelligence*, pages 3449–3468, Pune, India.

[Kulkarni and Pedersen, 2005] Kulkarni, A. and Pedersen, T. (2005). SenseClusters: unsupervised clustering and labeling of similar contexts. In *Proceedings of the Association for Computational Linguistics Interactive poster and demonstration sessions*, pages 105–108, Ann Arbor, MI.

[Leacock and Chodorow, 1998] Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

[Leacock et al., 1998] Leacock, C., Miller, G., and Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

[Leacock et al., 1993] Leacock, C., Towell, G., and Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Proceedings of the Workshop on Human Language Technology at the Annual Meeting of the Association for Computational Linguistics*, pages 260–265, Columbus, Ohio.

[Lee and Ng, 2002] Lee, Y. and Ng, H. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 41–48, Morristown, NJ.

[Leroy and Rindflesch, 2004] Leroy, G. and Rindflesch, T. (2004). Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a naıve bayes classifier. In *Proceedings of the 11th World Congress on Medical Informatics*, pages 381–385, San Fransico, CA.

[Leroy and Rindflesch, 2005] Leroy, G. and Rindflesch, T. (2005). Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *International Journal of Medical Informatics*, 74(7-8):573–85.

[Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Canada.

[Lin, 1997] Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Madrid, Spain.

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

[McCarthy, 1997] McCarthy, D. (1997). Word sense disambiguation for acquisition of selectional preferences. In *Proceedings of the Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications at the Annual Meeting for the Association of Computational Linguistics*, pages 52–61, Madrid, Spain.

[McInnes et al., 2009] McInnes, B., Pedersen, T., and Pakhomov, S. (2009). UMLS-Interface and UMLS-Similarity : Open source software for measuring paths and semantic similarity. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, San Fransico, CA.

[McRoy, 1992] McRoy, S. (1992). Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

[Miller and Charles, 1991] Miller, G. and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

[Mohammad and Hirst, 2006] Mohammad, S. and Hirst, G. (2006). Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 121–128, Trento, Italy.

[Mohammad and Pedersen, 2004] Mohammad, S. and Pedersen, T. (2004). Combining lexical and syntactic features for supervised word sense disambiguation. In *Proceedings of the Eighth Conference on Natural Language Learning at the Meeting of the Human Language Technology and the North American Chapter of the Association for Computational Linguistics*, pages 25–32, Boston, MA.

[Mooney, 1996] Mooney, R. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82–91, Pittsburg, PA.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

[Ng, 1997] Ng, H. (1997). Exemplar-based word sense disambiguation: Some recent improvements. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 208–213, Providence, RI.

[Ng and Lee, 1996] Ng, H. and Lee, H. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 16th Annual Meeting of the Association for Computational Linguistics*, pages 40–47, Santa Cruz, CA.

[Nguyen and Al-Mubaid, 2006] Nguyen, H. and Al-Mubaid, H. (2006). New ontology-based semantic similarity measure for the biomedical domain. In *Proceedings of the IEEE International Conference on Granular Computing*, pages 623–628, Atlanta, GA.

[Patwardhan, 2003] Patwardhan, S. (2003). Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. *Master's thesis, University of Minnesota, Duluth.*

[Patwardhan and Pedersen, 2006] Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.

[Pedersen, 2000] Pedersen, T. (2000). A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 63–69, Seattle, WA.

[Pedersen, 2001] Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Annual Meeting of the North American Chapter Of The Association For Computational Linguistics)*, pages 1–8, Pittsburgh, PA.

[Pedersen et al., 2005] Pedersen, T., Banerjee, S., and Patwardhan, S. (2005). Maximizing semantic relatedness to perform word sense disambiguation. *Supercomputing Institute Research Report UMSI*, 25.

[Pedersen and Bruce, 1997] Pedersen, T. and Bruce, R. (1997). Distinguishing word senses in untagged text. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, volume 2, pages 197–20, Providence, RI.

[Pedersen and Bruce, 1998] Pedersen, T. and Bruce, R. (1998). Knowledge lean word-sense disambiguation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 800–805, Madison, WI.

[Pedersen et al., 2007] Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

[Pedersen et al., 2004] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity - Measuring the relatedness of concepts. In *The Annual Meeting of the Human Language Techonology and North American Association of Computational Linguistics: Demonstration Papers*, pages 38–41, Boston, Massachusetts, USA.

[Purandare and Pedersen, 2004] Purandare, A. and Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Natural Language Learning*, pages 41–48, Boston, MA.

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

[Resnik, 1993] Resnik, P. (1993). Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 278–283, Plainsboro, NJ.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

[Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Savova et al., 2008] Savova, G., Coden, A., Sominsky, I., Johnson, R., Ogren, P., Groen, P., and Chute, C. (2008). Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics*, 41(6):1088–1100.

[Schütze, 1992] Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the ACM/IEEE Conference on Supercomputing*, pages 787–796, Minneapolis, MN.

[Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

[Stevenson et al., 2008] Stevenson, M., Guo, Y., Gaizauskas, R., and Martinez, D. (2008). Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):11.

[Stevenson and Wilks, 2001] Stevenson, M. and Wilks, Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

[Sussna, 1993] Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*, pages 67–74, Washington, DC.

[Wagstaff and Cardie, 2000] Wagstaff, K. and Cardie, C. (2000). Clustering with instance-level constraints. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 1097–1097, Austin, TX.

[Weaver, 1949] Weaver, W. (1949). Translation. *Mimeographed*, pages 15–23.

[Witten and Frank, 1999] Witten, I. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

[Wu and Palmer, 1994] Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Meeting of Association of Computational Linguistics*, pages 133–138, Las Cruces, NM.

[Yarowksy and Florian, 2003] Yarowksy, D. and Florian, R. (2003). Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(04):293–310.

[Yarowsky, 1992] Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th Meeting of the Association of Computational Linguistics*, pages 454–460, Newark, NJ.

[Yarowsky, 1993] Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the Workshop on Human Language Technology at the Meeting for the Association of Computational Linguistics*, pages 266–271, Columbus, Ohio.

[Yarowsky, 1995] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Meeting on Association for Computational Linguistics*, pages 189–196, Cambridge, MA.

Appendix A

Similarity Measures

Semantic similarity and relatedness measures assign a score to how similar or related two concepts are to each other. Semantic relatedness is a more general form of semantic similarity. For example, *foot* and *pedal edema* are not similar but are related, where as *foot* and *hand* are both similar and related. Similarity measures use the is-a relations in a hierarchy. There are two types: path-based and IC based. Path based measures require the shortest path between two concepts using an is-a hierarchy. IC based measures are classified as similarity measures because they require obtaining the least common subsumer (LCS) of the two concepts which is determined based on the is-a links. Relatedness uses more than that, and it can be path based including is-a relations but it includes addition information such as part-of relations and definitions.

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. The semantic similarity and most of the relatedness measures require the use of relation information such as the *is-a* relations in a lexical database. The most commonly used lexical database for general English is WordNet. WordNet is a machine readable dictionary whose words are organized into concepts that are connected together through a variety of relations. A more in depth discussion of WordNet can be found in Section 2.3.1.

In this chapter, we discuss 14 similarity and relatedness measures that have been used in WSD. We then discuss a comparative analysis of these measures to human analysis.

We discuss the path-based measures proposed by :

- [Rada et al., 1989]
- [Sussna, 1993]
- [Leacock and Chodorow, 1998]
- [Altintas et al., 2005]
- [Wu and Palmer, 1994]
- [Agirre and Rigau, 1996]

The IC measures proposed by:

- [Resnik, 1995]
- [Jiang and Conrath, 1997]
- [Lin, 1997]

The relatedness measures proposed by :

- [Lesk, 1986]
- [Cowie et al., 1992]
- [Hirst and St-Onge, 1998].
- [Patwardhan, 2003]

The measure proposed by [Patwardhan, 2003] is classified as a vector measure even though it is a similarity measure due to its algorithm. A more detailed discussion of this measure is in Section 2.2.3.

A.1 Path-based Similarity Measures

The Path measure is the reciprocal of the number of nodes between two conceptsc_1 and c_2.

[Rada et al., 1989] introduce the measure conceptual distance. Conceptual distance is calculated as the shortest path between two concepts in a database. [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 database (D) as defined in Equation A.1.

$$\text{sim}_{lch}(c_1, c_2) = -\log \frac{\text{minpath}(c_1, c_2)}{2 * D} \tag{A.1}$$

[Altintas et al., 2005] modified Leacock and Chodorow's 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. This measure is defined in Equation A.2. The $SpecFactor$ is calculated by taking the quotient of the depth of the concept and the cluster depth which in turn is defined as the deepest node in the cluster. The $LenFactor$ is the modification of Leacock and Chodorow's measure. It is the shortest path between the two concepts divided by twice the depth of the taxonomy.

$$\text{sim}_{altintas}(c_1, c_2) = \frac{1}{1 + \text{SpecFactor}c_1, c_2 + \text{LenFactor}c_1, c_2} \tag{A.2}$$

[Wu and Palmer, 1994] (sim_{wup}) introduce a measure that takes into consideration the depth of two concepts in a database. 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 A.3.

$$\text{sim}_{wup}(c_1, c_2) = \frac{2 * \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \tag{A.3}$$

[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 A.4 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.

$$\text{sim}_{cd}(c_1, c_2) = \frac{\sum_{i=1}^{m} \text{nhyp}^{i^{0.20}}}{\text{descendants}_{c_1}} \tag{A.4}$$

A.2 Information Content Similarity Measures

Information content (IC) measures the specificity of a concept in a lexical database. 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 corpus 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 A.5. As stated above the LCS is the most specific concept two concepts share as an ancestor.

$$\text{sim}_{res} = \text{IC}(\text{lcs}(c_1, c_2) = -\log(P(\text{lcs}(c_1, c_2))) \tag{A.5}$$

[Jiang and Conrath, 1997] and [Lin, 1997] extended Resnik's information content measure. Jiang and Conrath modified it to include the length of the path between the two concepts defining it as a distance measure. This measure is modified in WordNet::Similarity to return a similarity score by taking the distance measure reciprocal as seen in Equation A.6. Lin modified it to include the individual information x content of the two concepts as seen in Equation A.7.

$$\text{sim}_{jcn} = \frac{1}{\text{IC}(c_1) + \text{IC}(c_2) - 2 * \text{IC}(\text{lcs}(c_1, c_2))} \tag{A.6}$$

$$\text{sim}_{lin} = \frac{2 * \text{IC}(\text{lcs}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)} \tag{A.7}$$

A.3 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] introduces 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 et al., 1992] alleviate the first limitations by incorporating simulated annealing. They use the simulated annealing optimization algorithm to approximate the results of calculating all possible combinations of senses. Therefore, common words between the possible glosses of the concept 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 in WordNet but also the gloss of the related concepts. [Patwardhan, 2003] extends this approach further by including the glosses of the related concepts. This inclusion of "friends of friends" information alleviates the sparseness and does not require the exact matching of words. For example, *asprin* and $Ibprofren$ may not occur together in the same context very often but they occur with the word *headache*. The measure proposed by Patwardhan incorporates this type of information.

[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 A.8. The relations used to determine the path can be either is-a relations or part-of relations. [Budanitsky and Hirst, 2001] and [Pedersen et al., 2005] set C equal to eight and k equal to one this measure.

$$\text{sim}_{hirst}(c_1, c_2) = C - \text{pathLength}(c_1, c_2) - (k * \text{the number of direction changes}) \quad \text{(A.8)}$$

A.4 Comparative Analysis of Semantic and Relatedness Measures

In this section, we discuss the comparative analysis of the semantic similarity and relatedness measures that have been conducted by researchers. First, we describe the data that was used to conduct the analysis and second, we discuss the analysis itself.

A.4.1 Data

The data used to evaluate the various similarity and relatedness measures are the datasets reported by [Rubenstein and Goodenough, 1965], [Miller and Charles, 1991] and [Pedersen et al., 2007].

Rubenstein and Goodenough dataset contains 65 general english word pairs whose semantic relatedness was determined by 51 human subjects. The semantic relatedness of each term pair was annotated based on a four point scale : (4.0) practically synonymous, (3.0) related, (2.0) marginally related and (1.0) unrelated.

Miller and Charles (1991) used 30 out of the 65 word pairs and obtained their similarity judgements from 38 subjects using the same scale. The 30 word pairs that were chosen consisted of 10 pairs that were annotated to be very related, 10 that were somewhat related and 10 that were not related.

Pedersen, et. al.'s dataset contains 30 medical term pairs whose semantic similarity was determined by nine medical coders, who we refer to as Coders, and three physicians, who refer to as Physicians, from the Mayo Clinic. The semantic similarity of each term pair was annotated based on the same four point scale used by Rubenstein and Goodenough.

A.4.2 Analysis

Table A.1 shows the semantic and relatedness measures used by researchers. 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.

Table A.1 shows the correlation results between the similarity measures and human judgements reported by [Budanitsky and Hirst, 2006]

Table A.1: Analysis of Semantic and Relatedness Measures (Correlation)

		Budanitsky and Hirst, 2006		Patwardhan, and Pedersen, 2006		Pedersen, et. al. 2007	Caviedes and Cimino, 2004
		M&C	R&G	M&C	R&G	clinical	biomedical
Path-based Similarity Measures	[Rada et al., 1989]						0.60-0.70
	[Wu and Palmer, 1994]						
	[Leacock and Chodorow, 1998]	0.82	0.84	0.74	0.77	0.47	
	[Hirst and St-Onge, 1998]	0.74	0.79				
Information Content Similarity Measures	[Resnik, 1995]	0.74	0.78	0.72	0.72	0.55	
	[Jiang and Conrath, 1997]	0.85	0.78	0.73	0.75	0.55	
	[Lin, 1997]	0.83	0.82	0.70	0.72	0.69	
Relatedness Vector Measure	[Banerjee and Pedersen, 2003]			0.91	0.90		
	[Patwardhan, 2003]			0.81	0.83	0.76	

and [Patwardhan and Pedersen, 2006] using the Rubenstein & Goodenough (R&G) and Miller & Charles (M&C) data, and [Pedersen et al., 2007] and [Caviedes and Cimino, 2004] using a clinical and biomedical dataset respectively. Information about the datasets is described in Section 2.6. The authors all analyze their results using the Spearman's Rank Correlation Coefficient

[Budanitsky and Hirst, 2006] evaluate the path-based similarity measures proposed by [Hirst and St-Onge, 1998], [Leacock and Chodorow, 1998], and [Resnik, 1995] and the information content similarity measures proposed by [Jiang and Conrath, 1997] and [Lin, 1997]. The authors use the WordNet to obtain the path information and the Brown corpus to obtain the frequency information. They found that the measure proposed by Jiang and Conrath obtained the highest correlation score using the M&C data but the measure proposed by Lin obtained the highest correlation score using the R&G data.

[Patwardhan and Pedersen, 2006] evaluate the path-based similarity measures proposed by [Leacock and Chodorow, 1998] and [Resnik, 1995], the information content similarity measures proposed by [Jiang and Conrath, 1997] and [Lin, 1997], the relatedness measure proposed by [Banerjee and Pedersen, 2003] and the vector measure proposed by [Patwardhan, 2003] (described in Section 7.2.3. The authors use the WordNet to obtain the path and frequency information. They found that the measure proposed by Banerjee and Pedersen obtained the highest correlation score using both the M&C and R&G data.

[Pedersen et al., 2007] evaluate the measures proposed by [Leacock and Chodorow, 1998], and [Resnik, 1995] and the information content similarity measures proposed by [Jiang and Conrath, 1997] and [Lin, 1997] and the vector measure proposed by [Patwardhan, 2003] using their clinical dataset described above. The authors use SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terms) to obtain the path information and the Mayo Clinic corpus of clinical notes to obtain the frequency information. Pedersen, et. al. report the measure proposed by [Patwardhan, 2003] obtains a higher correlation score than the other measures.

[Caviedes and Cimino, 2004] apply the conceptual distance measure proposed by [Rada et al., 1989] on the UMLS. They modified the measure by using the PAR/CHD relations rather than RB/RN relations as originally proposed by Rada, et. al. They

authors use the MSH, ICD-9-CM, SNMI sources in the UMLS to obtain the path information. They analyze their results on a variety of different combinations of the data and sources. Their correlation results consistently fell between 0.60 and 0.80.

A.5 Conclusion

In this chapter, we introduced semantic similarity and relatedness measures. Similarity measures are more narrowly defined requiring the hierarchical information in the form of *is-a* relations while relatedness measures tend to be more flexible. These measures have been evaluated in of themselves although they have also played a role in language processing tasks from WSD described by [Patwardhan and Pedersen, 2006] to spelling correction described by [Budanitsky and Hirst, 2006].

Appendix B

UMLS Metathesaurus

B.1 Introduction

This report constitutes an overview of what was learned about the Unified Medical Language System (UMLS) and a justification of how it was incorporated it into the UMLS::Interface and UMLS::Similarity packages. The UMLS::Interface package is a Perl interface to the UMLS installed locally in a MySQL database. UMLS::Interface provides an API as well as a set of command line utility programs demonstrating how to use the API. The utility programs allows the user to explore the local UMLS installation. The UMLS::Similarity package is a suite of Perl modules that implement a number of semantic similarity measures to determine the similarity between concepts in the UMLS. This package is a porting of semantic similarity measures that have been developed for the general English lexical database WordNet [Fellbaum, 1998] and have been implemented via the WordNet-Similarity package [Pedersen et al., 2004]. This has previously been done in the biomedical domain using SNOMED-CT prior to its inclusion in the UMLS [Pedersen et al., 2007]. The current goal is to allow for the semantic similarity to be taken between concepts in SNOMED-CT existing in the UMLS as well as the other UMLS terminologies. To do this, the UMLS framework was explored to gain an understanding of its structure.

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 a 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.

The main focus of this chapter is the Metathesaurus which contains the relations between the concepts in the UMLS. The remainder of this report is broken into five sections. The first section describes the relational tables included in the Metathesaurus. The second section poses specific questions that arose from analysis of the Metathesaurus and the hopefully the answers to those questions. The third section describes the semantic similarity measures developed using WordNet and ported to the UMLS. The fourth section discuss past work that have used semantic similarity measures on different segments of the Metathesaurus. The last section lays out the justification for using specific segments of the Metathesaurus.

B.2 Metathesaurus

The UMLS Metathesaurus contains a vocabulary database of biomedical and health related concepts. The concepts associated with words and terms are enumerated via Concept Unique Identifiers (CUIs). For example, in the UMLS release 2008AB, two possible senses of *cold* are C0009264 which has the preferred term *Cold Temperature* and C0009443 which has the preferred term *Common Cold* . The preferred term is the term assigned to the CUI for descriptive purposes. The preferred term is often put in parentheses next to its CUI in this section for clarity. Currently, the Metathesaurus contains approximately 1.5 million CUIs.

The concepts in the UMLS come from over 100 different knowledge sources that have been semi-automatically integrated into a single source. A concept within a specific terminology is called an Atom Unique Identifier (AUI). An AUI is a specific concept from a specific source. For example, the National Cancer Institute Thesaurus contains the AUI A12785313 which has the preferred term *Cold*, and the SNOMED Clinical Terms (SNOMED-CT) contains the AUI A3292554 which as the preferred term *Low Temperature*. The AUIs from the different knowledge sources are semi-automatically combined to form CUIs. So for example, A12785313 (Cold) from NCI and A3292554 (Low Temperature) from the SNOMED-CT are mapped to the CUI C0009264 (Cold Temperature)

In some sources, there exists relations between AUIs. For example, in NCI there exists an *is-a* relation between A12785313 (Cold) and A7574004 (Temperature). The merging of the AUIs from different sources creates relations between the CUIs. Since, A12785313 (Cold) maps to C0009264 (Cold Temperature) and A7574004 (Temperature) maps to C0039476 (Temperature), the relation between A12785313 and A7574004 can be mapped to the CUI level creating a relation between C0009264 and C0039476.

There exist two files in the Metathesaurus with the relation information: MRHIER and MRREL. MRHIER contains a complete representation of all of the hierarchies present in the source vocabularies. For each concept that is part of a hierarchy in the source vocabulary, MRHIER contains the complete concept to root path in that hierarchy. These are referred to as the AUI paths, where each CUI will have an entry (row) for each of its AUIs that is from the hierarchical sources. Each AUI path is specific

to a single source, so a single CUI may have multiple paths depending on how many sources it is associated with it. If there exist multiple inheritance there will exist an entry in MRHIER for each AUI path.

MRREL contains hierarchical and non-hierarchical relations between CUIs (rather than AUIs) in the Metathesaurus from all sources. The relationships included in MRREL are:

- PAR/CHD: parent/child
- RB/RN: broader/narrower than
- SY: source asserted synonymy
- RO: Has a relationship other than synonymous, narrower, or broader
- RL: The two concepts are similar or "alike".
- RQ: related and possibly synonymous
- SIB: sibling

In both the MRREL and MRHIER tables the RELA field provides a more precise definition of the relation if provided by the source vocabulary. Table B.1 shows the RELA relations associated with the RB/RN and PAR/CHD relations.

The RELA relation is between the AUIs rather than the CUIs because the "meaning" of some of these relations is source dependent. This means that relations such as *is-a* relations for example might be defined slightly differently in individual sources, and those differences are preserved when they are incorporated into the UMLS, leading to variations in the Metathesaurus. For example, in the source RXNORM the *is-a* relations in RELA is actually an *instance of* relation.

B.3 Questions

The purpose of this analysis of the UMLS is to determine how to go about porting semantic similarity measures from the WordNet-Similarity package [Pedersen et al., 2004] which uses the WordNet to the UMLS::Similarity package which uses the UMLS. This was previously done in the Semantic-Similarity package using the SNOMED-CT terminology [Pedersen et al., 2007]. The similarity measures rely on the path information between two concepts. There exists two different concept levels in the UMLS, AUIs

Table B.1: RELA Relations Associated with RB/RN and PAR/CHD Relations

Relation	RB/RN	PAR/CHD
inverse_isa/isa	x	x
has_tradename	x	
precise_ingredient_of	x	
mapped_from	x	
has_form	x	
has_part	x	
has_conceptual_part	x	
mapped_to	x	
has_lab_number	x	
inverse_was_a	x	
contained_in	x	
has_version	x	
has_branch		x
has_subtype		x
has_tributary		x
codesystem_of		x

reflected in the MRHIER table and CUIs reflected in the MRREL table. There are also two different types of relations parent/child reflected in the MRHIER table and parent/child and narrower/broader reflected in the MRREL.

This section poses and attempt to answer some of the questions that arose during the analysis of the UMLS Metathesaurus MRREL and MRHIER tables. The questions arose in the attempt to understand the similarities and difference between the tables to determine what relations and concepts to use in the UMLS::Similarity package. In answering these questions, the UMLS versions 2008AA, 2008AB and 2009AA were used.

By what criteria is a source and its relations included in MRHIER and MRREL tables?

One of the first questions that arose in the analysis of MRHIER and MRREL is by what criteria is a source included in MRHIER and more explicitly under what criteria

are some sources not included in MRHIER but are included in MRREL. This question was posed in order to determine what information was included in MRREL but not in MRHIER and vise versa.

MRHIER contains a complete representation of all of the hierarchies present in the source vocabularies. The working definition of a hierarchy in the UMLS according to the documentation is:

> Any source-asserted multi-level organization of a sources vocabulary [1]

The nature and purpose of these hierarchies may be different between vocabularies. When a source vocabulary is incorporated into the UMLS, the UMLS editors study the vocabulary and decides whether there is an explicit hierarchical structure. This determination depends more on the nature and intention of the relations rather than on the label given to them. Thus, not all *is-a* relations are considered hierarchical. The hierarchical vocabularies and the hierarchical relations are included in MRHIER and MRREL as PAR/CHD relations. The non-hierarchical *is-a* relations are represented as RB/RN (rather than PAR/CHD) relations in MRREL but are not included in MRHIER. Similarly, not all *part-of* relations are hierarchical.

For example, the MEDLINEPLUS, MTH, RXNORM, SRC, and VANDF source exist in the MRREL file for the Level0 + SNOMED-CT UMLS view but do not exist in MRHIER. The MTH source refers to the Metathesaurus indexers and is discussed below. SRC refers to linking of the different sources themselves. The following sections provide examples which would preclude the remaining sources from being part of MRHIER.

RXNORM : In RXNORM the is-a relations in RELA are actually "instance of" relations. For example, in MRREL there exists a RB/RN relation between C0040865 and C0773689 as seen below:

```
C0040865|A10490134|SCUI|RN|C0773689|A10451351|SCUI|isa|R45060915||RXNORM|RXNORM|||N||
C0773689|A10451351|SCUI|RB|C0040865|A10490134|SCUI|inverse_isa|R45033417||RXNORM|RXNORM|||N||
```

While this might look like a traditional *is-a* relation it is actually an *instance-of* or perhaps *quantity-of* relation. Looking at what these CUIs represent, one appears to be a specific quantity (instance) of the more general class:

[1] http://www.nlm.nih.gov/research/umls/glossary.html

```
C0040865 [Triamcinolone Oral Paste]
C0773689 [Triamcinolone 0.001 MG/MG Oral Paste]
```

Therefore, it is denoted as an RN/RB and *isa/inverse-isa* relations in MRREL. So, while these relations are included in MRREL with RXNORM saying that these are an $is-a$ relation, MRHIER excludes these because they are not traditional *is-a* relations.

VANDF : A very similar sort of example can be found in VANDF. For example, in MRREL there exists a RB/RN relation between C0301532 and C1572214 as seen below:

```
C0301532|A12100360|AUI|RN|C1572214|A8452860|AUI|isa|R70576114||VANDF|VANDF|||N||
C1572214|A8452860|AUI|RB|C0301532|A12100360|AUI|inverse_isa|R70583316||VANDF|VANDF|||N||
```

Once again there is an RB/RN relations with an specified *is-a* relation in MRREL, which is excluded from MRHIER. Looking at what these CUIs consist of, this seems to be an *instance-of* relation, where there is a particular kind of vitamin as being an instance of multivitamins in general.

```
C0301532 [Multivitamin preparation]
C1572214 [MULTIVITAMINS FOR JOINT HEALTH HERBAL CAP/TAB]
```

MEDLINEPLUS : For MEDLINEPLUS, the relations appear to be topic headings (more like a library index perhaps) and not really *is-a* or any other sort of hierarchical relation. For example, "C1456592 [Child and Teen Health]" has 58 RN (narrower) relations. Some of them include:

- C0001593 [Adoption]
- C0008066 [Child Behavior Disorders]
- C0008071 [Child Development]
- C0008073 [Developmental Disabilities]
- C0008078 [Children's Health]
- C0011854 [Juvenile Diabetes]
- C0018273 [Growth Disorders]
- C0021294 [Premature Babies]
- C0025007 [Measles]

- C0032968 [Pregnancy]
- C0035235 [Respiratory Syncy Virus Infections]
- C0036370 [School Health]
- C0043167 [Whooping Cough]

These are all clearly topics within child and teen health, but they are not really hierarchical. In fact they seem to be a fairly traditional (and understandable) notion of narrower (ie narrower topics in this domain). Therefore, these are included in MRREL but not MRHIER.

To summarize, MRHIER contains only those sources whose hierarchical structure is explicitly defined and the nature and intention of the relations are considered to be hierarchical. The hierarchical vocabularies and the hierarchical relations are included in both MRHIER and MRREL as PAR/CHD relations. The non-hierarchical *is-a* relations are represented as RB/RN relations in MRREL but are not included in MRHIER.

Why does the UWDA source contain *part-of*, *has-part* and *tributary-of* relations in MRHIER and as PAR/CHD relations in MRREL:

The only source that contributes *part-of*, *has-part* and *tributary-of* relations to MRHIER is the University of Washington Digital Anatomist (UWDA) source. In MRHIER, there are 288,011 relations that seem to be made up of *part-of* relations from UWDA. This is the only source that contributes paths to MRHIER based on *part-of* relations, although it is not the only source in which part-of relations are found. For example, MRREL contains 47,505 part-of relations from SNOMED-CT that are not included in MRHIER.

The reason is because UWDA explicitly defines a hierarchy based on *part-of* relationships while SNOMED-CT does not. Even though SNOMED-CT also has *part-of* relationships, they are not intended to form a distinct hierarchy. This is also the case for the *has- part* and *tributary-of* relations. The hierarchical relationships in UWDA are represented as both PAR/CHD and RB/RN relationships. This duplication is redundant and is an exception due to some past processes that are no longer required.

The redundant RB/RN relationships can be ignored.

How do the relations in MRREL correspond to the relations in MRHIER?

All the PAR/CHD relations in MRREL come directly from MRHIER. Although, it is not possible to generate MRHIER from MRREL. MRHIER represents the full path-to-root from the sources whereas MRREL only represents pairwise relationships. While for most sources, the full path-to-root is a transitive closure of the pairwise PAR/CHD relationships, this does not hold true for some sources. One example is MSH. One MSH descriptor can have different children depending on its tree position [2] . In the documentation, it states:

> A classic example is the case of 'Accidents,' which has as a narrower term, 'Accident Prevention.' Clearly, 'Accident Prevention' is not part of 'Accidents' nor is it included in the class of 'Accidents.' ... A search for documents about accidents in MEDLINE should find documents about accident prevention. It is the relationship of "aboutness" that is fundamental to a hierarchy in a thesaurus used for document retrieval. The relationships of part/whole and class/subclass have a role in subject retrieval because if a subject is about a subclass, then it is also likely about the class. Further analysis of this notion of subject or aboutness might provide us with rules for assigning if not using a hierarchy in document retrieval (Maron, 1977; Harper, 1989) ...Arranging material hierarchically with these criteria should result in the placement of a descriptor in more than one hierarchy. [3]

The pairwise PAR/CHD relationships in MRREL do not carry tree position information and so cannot be used to derive the full path-to-root for MSH. On the other hand, it is possible to generate all PAR/CHD relationships in MRREL from MRHIER. There exist a few other other sources like MSH: AIR, OMS, SNM and USPMG.

What are MTH relations?

[2] discussion on MSH hierarchies can be found here: http://www.nlm.nih.gov/mesh/meshrels.html

[3] http://www.nlm.nih.gov/mesh/meshrels.html Section "HIERARCHICAL RELATIONSHIPS:TREES, SUBSUMPTION, AND MESH IN DOCUMENT RETRIEVAL"

An MTH relation is a relation that has been defined by the UMLS editors themselves rather than explicitly from a source. A main difference between MTH relations and relations from other sources are that these relations are between CUIs rather than AUIs. MTH relations in MRREL make up less than 3% of all relations. Their origins can be grouped into 3 categories.

The first category are those MTH relations that are derived from source vocabularies (historic). In the early years of Metathesaurus development, some source-derived relations were 'cloned' into UMLS CUI-level relations and were attributed to the Metathesaurus (SAB=MTH). This was prior to the push for source transparency where the AUI information was not included in the MRREL table. This happened when CUI1 REL CUI2 originated from two different sources, there was only one SAB column so the MTH label would be put in the SAB column rather than the picking one of the two different sources. These can be considered legacy relations and this practice is no longer continued.

The second category are those added by UMLS editors during editing. There are rules in UMLS editing that encourage editors to add CUI-level relations in certain situations. First, when some atoms are split out from a concept to form a new concept, a CUI-level relation is added to characterize the link between the old and the new concepts. Second, when genuine ambiguity prevents the merging of two identical strings into the same concept, editors create a CUI-level relation between the two concepts containing the ambiguous strings. These relations are added primarily to aid internal quality assurance, identify missed synonymy, create links to orphan concepts and facilitate future editing.

The third category are those implied from source information but relation not explicitly stated by a source vocabulary. There exists abut 1,000 of these types of relations. These are very specific cases in which implied relations are added algorithmically during the processing of a source vocabulary. They are the only MTH relations with non-null RELA values. One example is the RELA='exhibits' relations. In the processing of GO, if a GO term (e.g. GO:0019104 DNA glycosylase) is encountered that is the same as a term from another vocabulary (e.g. D045647 DNA Glycosylase from MSH) inside an existing UMLS concept with STY="Enzyme", it is known that the GO term actually

refers to the enzyme activity rather than the enzyme itself and the two are not synonymous. Therefore, a CUI-level 'exhibits' relation is added between the two to prevent them from merging, so that editors do not have to tease them apart manually. Since this relation is not asserted by GO it is attributed to MTH.

The first two groups form the majority of MTH relations. The first group of course are declining since the practice is no longer in place. As of right now though, there is no way to distinguish between the first two groups. Many of these relations are created primarily to facilitate editing and quality assurance, but they also provide valuable information for UMLS users as well and so they are included in the release files. Due to the ad hoc and varied manner in which these relations are created, they are not expected to form any complete network or hierarchy. It is also not surprising that they sometimes overlap (at the CUI level) with relations asserted by some source vocabularies.

There are quite a few duplicate relations which may be explained by the legacy relations. The count is 87,784 them. Meaning there exists an RB/RN relation tagged MTH and a RB/RN relation tagged by another source. For example:

```
C0000120|A1384345|AUI|PAR|C0020344|A1393900|AUI||R05221343||AOD|AOD|||N||
C0000120|A1384345|AUI|RB|C0020344|A1393900|AUI||R00695462||AOD|AOD|||N||
C0000120||CUI|RB|C0020344||CUI||R02762164||MTH|MTH||N|N||
C0020344|A1393900|AUI|CHD|C0000120|A1384345|AUI||R05213318||AOD|AOD|||N||
C0020344|A1393900|AUI|RN|C0000120|A1384345|AUI||R00720907||AOD|AOD|||N||
C0020344||CUI|RN|C0000120||CUI||R02958009||MTH|MTH||N|N||
```

Here are two CUIs (C0000120 and C0020344) that are connected through an RB/RN and CHD/PAR relation from AOD. And then again connected with an RB/RN relation through MTH.

Below is another example that may come from the legacy relations. CUI1 and CUI2 have a PAR/CHD relation from the sources in which they have in common (NCBI, SCTSPA and SNOMED-CT). But they also have been given an RB/RN relation by the Metathesaurus editors.

```
CUI1: C0678246 (Pfiesteria, SAI) CSP NCBI SCTSPA SNOMED-CT
CUI2: C0522430 (Pfiesteria piscicida (organismo)) MSH NCBI SCTSPA SNOMED-CT
REL: CHD (SNOMED-CT) (NCBI) (SCTSPA)
```

```
CUI1: C0678246 (Pfiesteria, SAI) CSP NCBI SCTSPA SNOMED-CT
CUI2: C0522430 (Pfiesteria piscicida (organismo)) MSH NCBI SCTSPA SNOMED-CT
REL: RN (MTH)
```

Here are the raw entries from MRREL for these two CUIs:

```
C0522430|A1047397|SCUI|PAR|C0678246|A1308595|SCUI||R70369285||NCBI|NCBI|||N||
C0678246|A1308595|SCUI|CHD|C0522430|A1047397|SCUI||R70100922||NCBI|NCBI|||N||

C0678246|A3300064|SCUI|CHD|C0522430|A3183937|SCUI|isa|R20021726|7684027|
  SNOMED-CT|SNOMED-CT|0|Y|N||
C0522430|A3183937|SCUI|PAR|C0678246|A3300064|SCUI|inverse_isa|R20490719|
  7684027|SNOMED-CT|SNOMED-CT|0|N|N||

C0678246|A7036177|SCUI|CHD|C0522430|A6998242|SCUI|isa|R67142517|7684027|SCTSPA|
  SCTSPA|0|Y|N||
C0522430|A6998242|SCUI|PAR|C0678246|A7036177|SCUI|inverse_isa|R66747060|7684027|
  SCTSPA|SCTSPA|0|N|N||

C0522430||CUI|RB|C0678246||CUI||R02778595||MTH|MTH||N|N||
C0678246||CUI|RN|C0522430||CUI||R02902682||MTH|MTH||N|N||
```

What and why are there orphan relations (ie nodes that do not have a parent or broader than relation)?

There are a number of CUIs in the Metathesaurus that do not have a corresponding PAR or RB relation. Normally, PAR/CHD and RB/RN relations come as a pair (except for the root nodes). There are a number of exceptions though. For example, Table B.2, shows the number of CUIs with an RN relation but do not have an RB relation.

It seems that RXNORM, MTH, and MSH are the major contributors to this "orphan" issue, which is quite interesting although GO and SNOMED-CT are also fairly significant contributors.

One possible theory is that, for RXNORM and VANDF, this is because the RN/RB relation is really being used to express an *instance-of* relation which is more of a pairwise than a hierarchical relation. For others, such as CSP they could be upper level categories. For example, the four RB/RN orphan CUIs can be seen in Table B.3 along

Table B.2: The Number of CUIs With an RN Relation but Not an RB Relation

Source	CUIs
AOD	20
AOT	9
CSP	4
GO	2558
MEDLINEPLUS	17
MSH	10180
MTH	12285
MTHMST	260
NCI	1
NDFRT	6
PDQ	152
RXNORM	19587
SCTSPA	7
SNOMED-CT	1697
SRC	56
UWDA	5
VANDF	260

with their preferred term and all of their sources in the 2008AB version of the UMLS.

Table B.3: Four RB/RN Orphan CUIs from CSP

CUI	Preferred Term	Sources
C0025118	medicine (field)	AOT, CSP, MTH
C0178642	food science/technology	CSP, MTH
C0596159	behavioral/social science	CSP, MTH
C0872087	technology/technique	CSP

The RB/RN orphan CUIs in the NCI and the UWDA sources though do not seem to have the same feel. For example, the NCI the orphan CUI can be seen in Table B.4 and he RB/RN orphan CUIs for UWDA can be seen in Table B.5.

There are a number of CUIs with a CHD relation that do not have an PAR relation. Table B.6 shows the CUIs that do not have a PAR relation along with the preferred term associated with the CUI from the MRCONSO table and its source. Basically,

Table B.4: RB/RN Orphan CUI from NCI

CUI	Preferred Term	Sources
C0282279	Oceana (localizacin geogrfica)	MTH, NCI

Table B.5: Five RB/RN Orphan CUIs from UWDA

CUI	Preferred Term	Sources
C1179477	Anatomical entity	UWDA
C1179901	Part of perixosome	UWDA
C1179910	Anatomical entity template	UWDA
C1180113	Ring protein subunit	UWDA
C1181723	Cortical cell of adrenal gland	UWDA

there exists one CUI per source and the CUI is the top level source CUI.

To summarize, all the orphan CUIs are those that have an RN relation but not a corresponding RB relation. This may be the case because the RB/RN have so many different possible meanings, that in some cases (perhaps) it might make sense that there be orphans (like in the case of "instance of" relations...)

B.4 CUI Hierarchy

The advantage of using the MRHIER table is that it contains the complete concept to root path in a sources hierarchy which allows for the fast retrieval of path information for a given source. It also contains the full path-to-root of CUIs from sources whose path-to-root is not a transitive closure of the pairwise PAR/CHD relation such as MSH.

The main advantage of using MRREL is the flexibility that MRREL provides. MRREL allows the use of using just relations that exist in a single source or adding the RB/RN relations that are added by the UMLS editors. This has the potential of providing addition information that may be relevant. It also allows for more than a single source to be used. The semantic similarity could be calculated between concepts that exist in different sources which would greatly increase the amount of coverage.

Appendix III discusses the actual implementation of the UMLS::Interface and UMLS::Similarity packages. It also validate the functionality of the framework by reproducing the results the previous work described by [Caviedes and Cimino, 2004],

Table B.6: CUIs With a CHD Relation but Not a PAR Relation

CUI	Term	Source
C0391807	UWDA	UWDA
C0995203	NCBI	NCBI
C1137112	ICD-9-CM	ICD-9-CM
C1140091	AIR	AIR
C1140093	CSP	CSP
C1140145	ICPC	ICPC
C1140162	AOD	AOD
C1140180	MTHCH	MTHCH
C1140228	Clinical Classifications Categories	CCS
C1140233	MTHHH	MTHHH
C1368719	SNOMED Clinical Terms version 20070731	SCTSPA
C1371271	NDFRT	NDFRT
C1549098	HL7V2.5	HL7V2.5
C1553931	HL7V3.0	HL7V3.0
C1579327	USPMG	USPMG
C1704485	AOT	AOT

[Pedersen et al., 2007] and [Nguyen and Al-Mubaid, 2006].

Appendix C

Semantic Types

This appendix contains the list of semantic types that exist in the 2008AB version of the Unified Medical Language System (UMLS).

Table C.1: UMLS Semantic Types

Abbreviation	Full Form
acab	Acquired Abnormality
bacs	Biologically Active Substance
bdsu	Body Substance
biof	Biologic Function
bmod	Biomedical Occupation or Discipline
bpoc	Body Part, Organ, or Organ Component
carb	Carbohydrate
celf	Cell Function
diap	Diagnostic Procedure
dsyn	Disease or Syndrome
fndg	Finding
ftcn	Functional Concept
gora	Government or Regulatory Activity
hlca	Health Care Activity
humn	Human
idcn	Idea or Concept
inbe	Individual Behavior

Table C.2: UMLS Semantic Types (continued)

Abbreviation	Full Form
inpr	Intellectural Product
lang	Language
lbpr	Laboratory Procedure
lbtr	Laboratory or Test Result
lipd	Lipid
mamm	Mammal
medd	Medical Device
menp	Mental Process
mnob	Manufactured Object
mobd	Mental or Behavioral Dysfunction
moft	Molecular Function
neop	Neoplastic Process
npop	Natural Phenomenon or Process
orga	Organism Attribute
orgf	Organism Function
ortf	Organ or Tissue Function
patf	Pathologic Function
popg	Population Group
qlco	Qualitative Concept
qnco	Quantitative Concept
resa	Research Activity
sbst	Substance
socb	Social Behavior
sosy	Sign or Symptom
spco	Spatial Concept
tmco	Temporal Concept
topp	Therapeutic or Preventive Procedure

Appendix D

UMLS Semantic Relations

This appendix contains the list of semantic relations that exist in the 2008AB version of the Unified Medical Language System (UMLS).

Table D.1: 2008 AB UMLS Semantic Relations

is a	associated with	physically related to
part of	consists of	contains
connected to	interconnects	branch of
tributary of	ingredient of	spatially related to
location of	adjacent to	surrounds
traverses	functionally related to	affects
manages	treats	disrupts
complicates	interacts with	prevents
brings about	produces	causes
performs	associated with	functionally related to
carries out	exhibits	practices
occurs in	process of	users
manifestation of	indicates	result of
temporally related to	co occurs with	precedes
conceptually related to	evaluation of	degree of
analyzes	assesses effect of	measurement of
measures	diagnoses	property of
derivative of	developmental form of	method of
conceptual part of	issue in	

Appendix E

NLM-WSD Dataset

This appendix contains the target words in the NLM-WSD dataset and their corresponding CUIs in the 1999 version of the Unified Medical Language System (UMLS).

Table E.1: Possible CUIs for each Target Word in the NLM-WSD datset

target word	CUI	Preferred Term
adjustment	C0376209	Individual Adjustment
	C0456081	Adjustment Action
	C0683269	Psychological adjustment
association	C0004083	Mental association
	C0699792	Relationship by association
blood pressure	C0005823	Blood Pressure
	C0005824	Blood Pressure Determination
	C0428878	Arterial pressure
cold	C0009264	Cold Temperature
	C0009443	Common Cold
	C0024117	Chronic Obstructive Airway Disease
	C0010412	Cold Therapy
	C0234192	Cold Sensation
condition	C0348080	Condition
	C0009647	Conditioning (Psychology
culture	C0010453	Anthropological Culture
	C0430400	Laboratory culture
degree	C0449286	Degree < 1 >
	C0542560	Degree < 2 >
depression	C0011570	Mental Depression
	C0460137	Depression motion
determination	C0680730	Adjudication
	C0243075	Determination
discharge	C0012621	Discharge, Body Substance
	C0030685	Patient Discharge

Table E.2: Possible CUIs for each Target Word in the NLM-WSD datset (Cont.)

target word	CUI	Preferred Term
energy	C0424589	Vitality
	C0542479	Energy (physics)
evaluation	C0220825	Evaluation
	C0175637	Health evaluation
extraction	C0684295	Extraction
	C0185115	Extraction, NOS
failure	C0699796	Failure
	C0231174	Failure, NOS
fat	C0424612	Obese build
	C0015677	Fatty acid glycerol esters
fit	C0036572	Seizures
	C0424576	Fit and well
fluid	C0302908	Liquid substance, NOS
	C0444611	Fluid
frequency	C0439603	Frequencies
	C0042023	Increased frequency of micturition
ganglion	C0085648	Benign cystic mucinous tumour
	C0017067	Ganglia
glucose	C0017725	Glucose
	C0337438	Glucose measurement
growth	C0018270	Growth < 1 >
	C0220844	Growth < 2 >
immunosuppression	C0021079	Therapeutic immunosuppression
	C0021080	Natural immunosuppression
implantation	C0029976	Blastocyst Implantation, natural
	C0021107	Implantation procedure
inhibition	C0021467	Psychological inhibition
	C0021469	inhibition, physical
japanese	C0376247	Japanese language
	C0022342	Japanese Population
lead	C0023175	Lead
	C0373667	Lead measurement, quantitative
man	C0024554	Male
	C0025266	Men
	C0086418	Homo sapiens
mole	C0439189	mol
	C0026386	Mole the mammal
	C0349514	Benign melanocytic nevus of skin
mosaic	C0439750	Spatial Mosaic
	C0026578	Mosaicism
	C0700058	Mosaic
nutrition	C0392209	Nutrition
	C0028707	Science of nutrition
	C0600072	Feeding and dietary regimes
pathology	C0030664	Pathology
	C0677042	Pathology < 3 >
pressure	C0033095	Pressure- physical agent
	C0460139	Pressure - action
	C0234222	Baresthesia
radiation	C0034519	Electromagnetic Energy
	C0034618	Radiation therapy
reduction	C0441610	Reduction - action
	C0301630	Chemical Reduction

Table E.3: Possible CUIs for each Target Word in the NLM-WSD datset (Cont.)

target word	CUI	Preferred Term
repair	C0374711	Repair - action
	C0043240	Wound Healing
resistance	C0683598	Resistance < 1 >
	C0237834	Resistance < 2 >
scale	C0222045	Integumentary scale
	C0349674	Intellectual scale
	C0175659	Weight measurement scales
secretion	C0036537	Bodily secretions
	C0687157	Secretion
sensitivity	C0036667	Statistical sensitivity
	C0312418	Personality Sensitivity
	C0427965	Antimicrobial susceptibility
sex	C0009253	Coitus
	C0036862	Sex
	C0079399	Gender
single	C0087136	Unmarried
	C0205171	Singular
strains	C0080194	Muscle strain
	C0456178	Microbiology subtype strains
support	C0344211	Support
	C0183683	Support, NOS
surgery	C0038894	Surgery specialty
	C0600001	Surgery
transient	C0205374	Transitory
	C0040704	Transient Population Group
transport	C0005528	Biological Transport
	C0150390	Patient Transport
ultrasound	C0041618	Ultrasonography
	C0041621	Ultrasonic Shockwave
variation	C0042333	Genetic Variation
	C0205419	Variant
weight	C0043100	Weight
	C0005910	Body Weight
white	C0220938	White color
	C0007457	Caucasoid Race

Appendix F

Stoplist

This appendix contains the stoplist used in this dissertation.

Table F.1: K-CUI and A-CUI Stoplist

a	about	after	all	also	an	and	are
as	at	back	be	because	been	before	being
between	but	by	can	could	do	even	first
for	from	get	good	had	has	have	he
her	his	how	i	if	in	into	is
it	its	just	look	make	many	more	most
much	must	new	no	not	now	of	off
on	one	only	open	or	other	our	out
over	own	people	she	so	some	than	that
the	their	them	then	there	they	this	those
through	time	to	two	up	us	very	was
way	we	well	were	what	when	which	while
who	will	with	would	years	you	your	aacute
amp	ccaron	dollar	eacute	equo	hellip	iacute	icirc
ins	lsqb	mdash	ndash	oacute	oslash	rcaron	rsqb
scaron	uuml	yacute	zcaron				

Appendix G

A-CUI Result Tables

This appendix contains the A-CUI results of all of the experiments discussed in Chapter 5.

Table G.1: UMLS CUI Definition Results using the Euclidean Distance

target word	Random	CUI		PAR		CHD		SIB		SY	
	Baseline	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.39	0.14	0.46	0.37	0.31	0.15	**0.59**	0.14	0.39	0.14
blood pressure	0.38	**0.49**	0.32	0.34	0.07	0.19	0.16	0.43	0.19	**0.49**	0.32
cold	0.14	0.13	0.05	0.14	0.15	0.22	0.41	0.03	**0.47**	0.13	0.05
condition	0.54	0.28	0.02	0.88	0.36	0.89	0.03	**0.98**	0.02	0.28	0.02
culture	0.44	0.46	0.12	0.70	0.12	**0.89**	0.14	**0.89**	0.12	0.46	0.12
degree	0.49	0.11	**0.97**	0.20	0.83	0.58	0.94	0.31	0.82	0.11	**0.97**
depression	0.46	0.75	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.75	**1.00**
determination	0.44	**0.97**	0.00	**0.97**	0.00	0.28	**0.97**	0.52	0.00	**0.97**	0.00
discharge	0.40	**0.89**	0.16	0.79	0.03	0.76	0.01	0.59	0.07	**0.89**	0.16
energy	0.44	**0.93**	0.09	0.91	0.06	0.50	0.18	0.29	0.41	**0.93**	0.09
evaluation	0.52	0.44	0.50	0.42	**0.60**	0.44	0.53	0.43	0.50	0.56	0.54
extraction	0.43	0.26	0.06	0.36	0.06	**0.51**	0.06	**0.51**	0.06	0.26	0.06
failure	0.41	0.72	**0.86**	0.72	**0.86**	0.72	**0.86**	0.72	**0.86**	0.72	**0.86**
fat	0.51	0.75	0.48	0.42	0.23	0.44	0.48	0.03	**0.88**	0.75	0.48
fit	0.56	0.89	0.06	**1.00**	0.06	**1.00**	0.06	**1.00**	0.06	0.89	0.06
fluid	0.48	0.90	0.07	0.69	0.23	0.35	0.43	0.12	**0.99**	0.90	0.07
frequency	0.53	0.36	0.01	0.34	0.00	0.20	**0.91**	0.36	0.01	0.26	0.00
ganglion	0.52	**0.93**	0.10	**0.93**	0.11	0.10	0.17	0.07	0.17	**0.93**	0.10
glucose	0.54	0.42	0.84	0.22	0.47	0.16	0.85	0.09	**0.86**	0.42	0.84
growth	0.61	0.42	0.37	0.51	0.37	0.62	0.37	**0.63**	0.37	0.42	0.37
immunosuppression	0.48	**0.60**	0.57	0.39	0.58	0.53	0.59	0.45	0.59	**0.60**	0.57
implantation	0.49	0.46	0.19	**0.83**	0.00	0.19	**0.83**	**0.83**	0.78	0.46	0.19
inhibition	0.53	0.48	0.01	0.26	0.34	0.28	0.25	**0.99**	0.01	0.48	0.01
japanese	0.56	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	0.00	**0.94**
lead	0.21	**0.59**	0.31	0.21	0.48	0.28	0.52	0.07	0.45	**0.59**	0.31
man	0.26	0.14	0.22	0.37	0.27	**0.65**	0.33	0.63	0.40	0.14	0.18
mole	0.39	0.51	**0.99**	0.18	**0.99**	0.80	**0.99**	0.51	**0.99**	0.51	**0.99**
mosaic	0.37	0.44	0.12	0.46	0.35	0.48	0.20	0.29	**0.56**	0.44	0.12
nutrition	0.42	**0.42**	0.16	0.30	0.19	0.31	0.17	0.31	0.15	**0.42**	0.16
pathology	0.45	0.48	0.13	0.46	0.78	**0.82**	0.81	0.14	0.64	0.48	0.13
pressure	0.28	0.74	0.71	0.52	0.92	0.65	**1.00**	0.28	0.54	0.74	0.71
radiation	0.52	0.45	**0.60**	0.34	**0.60**	0.39	**0.60**	0.33	**0.60**	0.45	**0.60**
reduction	0.36	0.27	**0.82**	0.27	0.73	0.27	0.73	0.36	0.64	0.27	**0.82**
repair	0.41	0.43	0.34	0.46	0.66	0.50	**0.75**	0.74	0.35	0.43	0.34
resistance	0.67	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
scale	0.32	0.65	**1.00**	0.68	0.12	0.23	0.00	0.65	**1.00**	0.65	**1.00**
secretion	0.53	0.45	0.01	0.72	0.01	**0.99**	0.01	**0.99**	0.01	0.45	0.01
sensitivity	0.31	0.24	0.02	0.24	0.02	**0.84**	0.02	0.39	0.02	0.24	0.02
sex	0.29	0.25	0.15	0.43	0.14	0.17	0.13	0.36	**0.58**	0.25	0.15
single	0.53	0.42	**0.98**	0.49	0.83	0.44	0.01	0.49	0.91	0.42	**0.98**
strains	0.49	**0.73**	0.01	**0.73**	0.01	0.51	0.01	**0.73**	0.01	**0.73**	0.01
support	0.80	0.40	**0.70**	0.20	**0.70**	0.20	**0.70**	0.40	**0.70**	0.40	**0.70**
surgery	0.50	0.16	0.63	0.38	0.88	0.21	0.88	**0.97**	0.72	0.16	0.63
transient	0.52	0.35	**0.98**	0.47	**0.98**	0.46	**0.98**	0.35	**0.98**	0.35	**0.98**
transport	0.53	0.52	**0.97**	0.12	0.96	0.01	0.96	0.01	0.95	0.52	**0.97**
ultrasound	0.43	0.33	**0.83**	0.25	**0.83**	0.16	0.00	0.16	**0.83**	0.33	**0.83**
variation	0.54	0.48	0.20	0.50	0.42	**0.80**	0.27	**0.80**	0.20	0.48	0.20
weight	0.51	0.55	0.49	0.42	0.53	0.45	0.45	0.43	**0.68**	0.55	0.49
white	0.49	**0.62**	0.48	0.51	0.49	0.46	0.47	0.49	0.56	**0.62**	0.48
Overall Accuracy	0.46	**0.50**	0.40	0.47	0.42	0.45	0.46	0.46	0.48	0.48	0.40

Table G.2: UMLS CUI Definition Results using the Cosine Measure

target word	Random	CUI		PAR		CHD		SIB		SY	
	Baseline	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.37	**0.66**	0.20	0.30	0.17	0.52	0.15	0.24	0.37	**0.66**
blood pressure	0.38	0.42	0.14	0.38	**0.51**	0.47	0.23	0.27	0.18	0.42	0.14
cold	0.14	0.12	0.00	0.12	**0.20**	0.04	0.02	0.18	0.13	0.12	0.00
condition	0.54	0.21	0.07	**0.55**	0.29	0.34	0.12	0.04	0.02	0.21	0.07
culture	0.44	0.42	0.29	0.28	0.41	0.16	**0.51**	0.14	0.26	0.42	0.29
degree	0.49	**0.06**	0.03	0.03	0.03	0.05	0.03	0.03	0.03	**0.06**	0.03
depression	0.46	0.91	0.91	0.78	0.89	**0.99**	0.95	0.85	0.85	0.91	0.91
determination	0.44	0.97	**1.00**	0.97	**1.00**	0.06	0.05	0.63	**1.00**	0.97	**1.00**
discharge	0.40	**0.92**	0.79	0.68	0.88	0.85	0.71	0.91	0.89	**0.92**	0.79
energy	0.44	0.97	**0.99**	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**	0.97	**0.99**
evaluation	0.52	0.44	0.55	0.46	0.52	0.49	0.52	**0.59**	0.54	0.58	0.54
extraction	0.43	0.10	0.05	0.11	0.06	0.09	0.08	**0.86**	0.15	0.10	0.05
failure	0.41	0.66	**0.79**	0.66	**0.79**	0.66	**0.79**	0.66	**0.79**	0.66	**0.79**
fat	0.51	0.85	0.14	0.86	0.14	**0.89**	0.15	0.84	0.18	0.85	0.14
fit	0.56	0.06	**0.11**	0.06	0.06	0.06	0.06	0.06	0.06	0.06	**0.11**
fluid	0.48	0.98	**1.00**	0.95	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.98	**1.00**
frequency	0.53	0.39	0.73	0.32	0.69	0.71	**0.77**	0.39	0.73	0.23	0.73
ganglion	0.52	**0.93**	0.66	**0.93**	0.52	**0.93**	0.62	**0.93**	0.22	**0.93**	0.66
glucose	0.54	0.83	0.19	0.82	0.19	0.25	0.15	**0.91**	0.49	0.83	0.19
growth	0.61	0.46	**0.63**	0.51	0.61	0.52	0.61	0.52	0.61	0.46	**0.63**
immunosuppression	0.48	0.59	**0.70**	0.45	0.65	0.59	0.67	0.55	0.60	0.59	**0.70**
implantation	0.49	**0.61**	0.34	0.37	0.28	0.60	0.40	0.37	0.58	**0.61**	0.34
inhibition	0.53	**0.49**	0.08	0.43	0.44	0.14	0.05	0.08	0.04	**0.49**	0.08
japanese	0.56	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	**0.86**	0.10	0.79	0.14	**0.86**	0.10	0.24	0.10	**0.86**	0.10
man	0.26	0.50	**0.53**	0.35	0.32	0.12	0.16	0.08	0.33	0.48	0.50
mole	0.39	0.86	**1.00**	0.87	0.99	0.86	**1.00**	0.86	**1.00**	0.86	**1.00**
mosaic	0.37	**0.61**	0.58	0.57	0.59	**0.61**	0.58	0.55	0.49	**0.61**	0.58
nutrition	0.42	0.21	0.28	0.21	**0.31**	0.17	0.21	0.18	0.20	0.21	0.28
pathology	0.45	0.31	0.33	0.55	0.39	0.23	0.21	**0.59**	0.34	0.31	0.33
pressure	0.28	0.77	0.73	0.59	0.57	**1.00**	0.98	0.92	0.48	0.77	0.73
radiation	0.52	0.59	0.61	**0.65**	0.50	0.60	0.51	0.61	0.60	0.59	0.61
reduction	0.36	0.27	0.82	0.73	0.73	0.27	0.82	**0.91**	0.82	0.27	0.82
repair	0.41	0.34	0.40	0.34	0.60	0.43	0.53	0.25	**0.62**	0.34	0.40
resistance	0.67	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
scale	0.32	0.69	0.80	0.74	**0.82**	0.45	0.00	0.69	0.80	0.69	0.80
secretion	0.53	0.39	**0.87**	0.34	0.86	0.45	0.03	0.42	0.27	0.39	**0.87**
sensitivity	0.31	0.24	**0.75**	0.24	**0.75**	0.12	0.51	0.22	0.04	0.24	**0.75**
sex	0.29	0.24	0.46	0.22	0.60	**0.68**	0.63	0.20	0.15	0.24	0.46
single	0.53	**0.41**	0.01	0.32	0.01	**0.41**	0.01	0.06	0.01	**0.41**	0.01
strains	0.49	**0.73**	0.13	**0.73**	0.13	0.48	0.61	**0.73**	0.13	**0.73**	0.13
support	**0.80**	**0.80**	**0.80**	**0.80**	0.60	0.70	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**
surgery	0.50	0.01	0.02	**0.03**	0.02	0.01	0.02	**0.03**	0.02	0.01	0.02
transient	0.52	0.33	0.01	**0.47**	0.02	0.33	0.01	0.33	0.01	0.33	0.01
transport	0.53	0.88	0.51	0.88	0.39	**0.96**	0.49	0.94	0.38	0.88	0.51
ultrasound	0.43	0.75	0.59	0.81	0.58	**0.83**	0.70	0.81	0.39	0.75	0.59
variation	0.54	**0.30**	0.21	0.23	0.20	0.25	0.25	0.21	0.20	**0.30**	0.21
weight	0.51	0.55	**0.64**	0.43	0.60	0.55	0.62	0.57	0.62	0.55	**0.64**
white	0.49	**0.59**	0.51	0.53	0.49	0.43	0.50	0.56	0.56	**0.59**	0.51
Overall Accuracy	0.46	**0.53**	0.50	0.51	0.50	0.49	0.45	0.50	0.45	0.53	0.50

Table G.3: UMLS CUI Definition Results using the Dice Coefficient

target word	Random	CUI		PAR		CHD		SIB		SY	
	Baseline	o1	o2	o1	o2	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.27	**0.69**	0.15	0.18	0.15	0.56	0.14	0.23	0.27	**0.69**
blood pressure	0.38	0.49	0.53	0.25	0.48	**0.55**	0.53	0.24	0.14	0.49	0.53
cold	0.14	0.11	0.01	0.11	0.15	0.07	0.05	0.51	**0.69**	0.11	0.01
condition	0.54	0.12	0.15	0.60	**0.71**	0.36	0.52	0.09	0.09	0.12	0.15
culture	0.44	0.12	0.11	0.10	0.10	0.12	**0.31**	0.11	0.09	0.12	0.11
degree	0.49	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
depression	0.46	0.93	**0.95**	0.91	0.92	0.92	**0.95**	0.92	**0.95**	0.93	**0.95**
determination	0.44	0.99	**1.00**	0.99	**1.00**	0.03	0.00	0.42	**1.00**	0.99	**1.00**
discharge	0.40	0.93	0.96	0.21	0.85	0.88	**0.97**	**0.97**	0.96	0.93	0.96
energy	0.44	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
evaluation	0.52	**0.54**	0.50	0.48	0.51	0.51	0.50	0.51	0.50	0.51	0.50
extraction	0.43	0.05	0.05	0.05	0.05	0.05	0.05	**0.91**	0.75	0.05	0.05
failure	0.41	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**
fat	0.51	0.95	0.93	**0.97**	0.93	**0.97**	0.93	**0.97**	0.93	0.00	0.93
fit	0.56	0.00	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**	0.00	**0.06**
fluid	0.48	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**1.00**
frequency	0.53	0.38	0.95	0.24	0.86	**0.97**	**0.97**	0.38	0.95	0.06	0.85
ganglion	0.52	0.93	**0.97**	0.93	0.95	0.93	**0.97**	0.93	0.34	0.93	**0.97**
glucose	0.54	0.89	0.89	0.89	0.87	0.59	0.55	**0.90**	0.87	0.89	0.89
growth	0.61	0.34	0.37	**0.38**	0.37	0.37	0.37	**0.38**	0.37	0.34	0.37
immunosuppression	0.48	0.52	0.54	0.51	0.55	0.56	0.52	0.53	**0.64**	0.52	0.54
implantation	0.49	0.70	**0.78**	0.29	0.30	0.77	0.63	0.44	0.76	0.70	**0.78**
inhibition	0.53	0.35	0.02	**0.72**	0.69	0.03	0.01	0.07	0.11	0.35	0.02
japanese	0.56	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	0.90	**0.93**	**0.93**	**0.93**	0.90	**0.93**	**0.93**	**0.93**	0.90	**0.93**
man	0.26	0.73	0.37	0.30	0.14	0.18	0.33	0.23	0.45	**0.75**	0.37
mole	0.39	0.98	**0.99**	**0.99**	**0.99**	0.98	**0.99**	0.98	**0.99**	0.98	**0.99**
mosaic	0.37	0.53	0.53	0.52	**0.55**	0.53	0.53	0.52	0.54	0.53	0.53
nutrition	0.42	0.13	0.19	0.22	0.22	0.12	0.18	0.22	**0.39**	0.13	0.19
pathology	0.45	0.15	0.25	0.74	**0.83**	0.29	0.71	0.56	0.18	0.15	0.25
pressure	0.28	0.98	0.98	0.75	0.98	**1.00**	0.97	**1.00**	0.97	0.98	0.98
radiation	0.52	0.57	0.58	0.58	0.58	0.56	0.57	**0.59**	0.57	0.57	0.58
reduction	0.36	0.36	**0.82**	**0.82**	**0.82**	0.36	**0.82**	**0.82**	**0.82**	0.36	**0.82**
repair	0.41	0.32	0.29	0.32	**0.35**	0.34	0.29	0.26	0.29	0.32	0.29
resistance	0.67	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
scale	0.32	0.88	**1.00**	0.80	0.98	0.22	0.02	0.88	**1.00**	0.88	**1.00**
secretion	0.53	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
sensitivity	0.31	**0.16**	0.02	**0.16**	0.02	0.02	0.02	0.02	0.02	**0.16**	0.02
sex	0.29	0.27	0.16	0.15	0.16	**0.86**	0.80	0.14	0.19	0.27	0.16
single	0.53	**0.15**	0.01	**0.15**	0.01	**0.15**	0.01	0.01	0.01	**0.15**	0.01
strains	0.49	**0.54**	0.14	**0.54**	0.14	0.02	0.04	**0.54**	0.14	**0.54**	0.14
support	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**
surgery	0.50	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**
transient	0.52	0.40	0.01	**0.60**	0.01	0.40	0.01	0.40	0.01	0.40	0.01
transport	0.53	**0.98**	**0.98**	0.97	**0.98**	0.97	0.97	0.96	0.96	**0.98**	**0.98**
ultrasound	0.43	0.80	0.73	0.80	0.73	**0.81**	0.72	0.74	0.73	0.80	0.73
variation	0.54	0.22	0.20	0.22	0.20	**0.26**	0.21	0.20	0.20	0.22	0.20
weight	0.51	0.55	0.57	**0.64**	0.57	0.30	0.42	0.51	0.57	0.55	0.57
white	0.49	0.58	0.48	0.50	**0.59**	0.52	0.49	0.39	0.49	0.58	0.48
Overall Accuracy	0.46	0.52	0.54	0.51	**0.55**	0.48	0.51	0.51	0.54	0.49	0.53

Table G.4: UMLS Preferred Term Results

target word	Random	Euclidean		Cosine		Dice	
	Baseline	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.32	0.19	0.37	0.14	0.40	**0.69**
blood pressure	0.38	0.36	0.48	0.42	**0.53**	0.31	**0.53**
cold	0.14	**0.58**	0.01	0.12	0.56	0.28	0.01
condition	0.54	0.42	**0.98**	0.21	0.46	0.10	0.15
culture	0.44	**0.56**	0.11	0.42	0.15	0.46	0.11
degree	0.49	0.28	**0.97**	0.06	0.03	0.15	0.03
depression	0.46	0.24	0.00	0.91	0.15	0.20	**0.95**
determination	0.44	0.97	0.00	0.97	**1.00**	0.99	**1.00**
discharge	0.40	0.35	0.01	0.92	0.83	0.40	**0.96**
energy	0.44	0.90	0.04	0.97	**0.99**	0.97	**0.99**
evaluation	0.52	**0.50**	**0.50**	0.44	**0.50**	0.45	**0.50**
extraction	0.43	0.42	0.06	0.10	**0.58**	0.19	0.05
failure	0.41	0.79	**0.86**	0.66	0.17	**0.86**	0.00
fat	0.51	0.78	**0.97**	0.85	0.47	0.81	0.93
fit	0.56	0.61	0.06	0.06	**1.00**	0.94	0.06
fluid	0.48	0.89	0.07	0.98	**1.00**	**1.00**	**1.00**
frequency	0.53	0.07	**0.99**	0.39	0.00	0.01	0.95
ganglion	0.52	0.93	0.89	0.93	0.07	0.93	**0.97**
glucose	0.54	0.16	0.87	0.83	0.09	0.09	**0.89**
growth	0.61	0.59	0.37	0.46	**0.63**	**0.63**	0.37
immunosuppression	0.48	0.53	0.58	**0.59**	0.57	0.51	0.54
implantation	0.49	0.60	0.17	0.61	**0.82**	0.59	0.78
inhibition	0.53	0.69	0.06	0.49	0.73	**0.77**	0.02
japanese	0.56	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	0.07	0.72	0.86	0.07	0.07	**0.93**
man	0.26	0.33	0.04	**0.50**	0.36	0.36	0.37
mole	0.39	0.01	**0.99**	0.86	0.92	0.00	0.00
mosaic	0.37	0.37	0.03	**0.61**	0.54	0.45	0.53
nutrition	0.42	0.28	**0.47**	0.21	0.28	0.22	0.19
pathology	0.45	0.51	0.14	0.31	**0.86**	0.70	0.25
pressure	0.28	0.47	0.12	0.77	0.14	0.29	**0.98**
radiation	0.52	0.43	0.60	0.59	**0.67**	0.41	0.58
reduction	0.36	0.27	**0.82**	0.27	**0.82**	0.36	**0.82**
repair	0.41	**0.51**	0.25	0.34	0.25	0.41	0.29
resistance	0.67	0.00	0.00	0.00	**1.00**	0.00	**1.00**
scale	0.32	0.17	0.00	0.69	0.14	0.14	**1.00**
secretion	0.53	0.17	0.01	0.39	**0.43**	0.04	0.01
sensitivity	0.31	0.24	0.02	0.24	**0.75**	0.16	0.02
sex	0.29	0.41	0.12	0.24	0.05	**0.42**	0.16
single	0.53	0.18	**0.98**	0.41	0.01	0.01	0.01
strains	0.49	**0.73**	0.01	**0.73**	0.13	0.54	0.14
support	**0.80**	0.20	0.30	**0.80**	0.20	0.20	**0.80**
surgery	0.50	**0.10**	0.07	0.01	0.02	0.02	0.02
transient	0.52	0.13	**0.98**	0.33	0.01	0.01	0.01
transport	0.53	0.44	0.02	0.88	0.09	0.30	**0.98**
ultrasound	0.43	0.26	**0.83**	0.75	0.16	0.17	0.73
variation	0.54	0.32	**0.80**	0.30	0.20	0.24	0.20
weight	0.51	**0.57**	0.45	0.55	0.55	0.55	**0.57**
white	0.49	0.53	0.46	**0.59**	0.31	0.50	0.48
Overall Accuracy	0.46	0.43	0.40	**0.53**	0.44	0.40	0.50

Table G.5: UMLS Associated Term Results

target word	Random	Euclidean		Cosine		Dice	
	Baseline	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.48	0.67	0.48	0.61	0.51	**0.68**
blood pressure	0.38	0.20	**0.39**	0.27	0.28	0.11	0.02
cold	0.14	0.78	0.01	0.49	0.36	0.08	**0.86**
condition	0.54	0.79	0.02	0.60	**0.97**	0.18	0.15
culture	0.44	0.50	0.11	**0.54**	0.11	0.43	0.11
degree	0.49	0.37	**0.97**	0.40	0.03	0.20	0.03
depression	0.46	0.78	0.04	**0.95**	0.79	0.88	**0.95**
determination	0.44	0.41	0.00	0.54	**0.92**	0.30	0.00
discharge	0.40	0.47	**0.99**	0.49	0.89	0.56	0.88
energy	0.44	0.78	**0.99**	0.72	0.07	0.83	0.98
evaluation	0.52	0.50	0.50	0.48	0.46	0.50	**0.60**
extraction	0.43	**0.40**	0.06	0.23	0.05	0.09	0.05
failure	0.41	0.72	**0.86**	0.79	0.41	0.83	0.83
fat	0.51	0.73	0.03	**0.79**	0.75	0.47	0.58
fit	0.56	0.83	0.06	0.83	0.72	0.89	**1.00**
fluid	0.48	0.35	**1.00**	0.36	0.01	0.76	0.92
frequency	0.53	0.64	**0.65**	0.18	0.06	0.02	0.02
ganglion	0.52	**0.93**	0.07	**0.93**	0.90	**0.93**	**0.93**
glucose	0.54	0.13	**0.91**	0.45	0.80	0.73	0.85
growth	0.61	0.36	0.37	0.39	**0.42**	0.38	0.37
immunosuppression	0.48	0.55	0.44	0.52	**0.59**	0.53	0.53
implantation	0.49	0.70	**0.86**	0.60	0.72	0.51	0.70
inhibition	0.53	0.56	0.02	0.81	**1.00**	0.94	0.99
japanese	0.56	0.35	0.91	0.46	0.58	0.06	**0.94**
lead	0.21	0.59	0.07	0.83	**0.93**	0.59	0.07
man	0.26	0.04	0.35	0.46	0.07	**0.47**	0.00
mole	0.39	0.36	0.00	0.36	**0.99**	0.32	**0.99**
mosaic	0.37	0.53	0.54	0.43	**0.60**	0.52	0.49
nutrition	0.42	0.31	0.31	0.33	0.38	0.33	**0.39**
pathology	0.45	0.30	0.14	0.25	**0.37**	0.15	0.16
pressure	0.28	0.32	0.78	0.12	0.02	0.47	**0.98**
radiation	0.52	0.38	0.40	0.42	**0.70**	0.43	0.67
reduction	0.36	0.45	**0.82**	0.45	**0.82**	0.55	**0.82**
repair	0.41	0.76	0.24	0.71	**0.78**	0.40	0.29
resistance	0.67	0.33	0.00	0.33	**1.00**	**1.00**	**1.00**
scale	0.32	**0.20**	0.00	0.12	0.03	0.14	0.02
secretion	0.53	0.18	0.01	**0.32**	0.05	0.01	0.01
sensitivity	0.31	**0.25**	0.02	0.12	0.14	0.12	0.02
sex	0.29	0.50	0.22	0.49	**0.80**	0.54	**0.80**
single	0.53	0.29	**0.99**	0.31	0.05	0.26	0.08
strains	0.49	0.48	0.01	0.58	**0.62**	0.42	0.35
support	0.80	0.30	0.50	0.70	0.60	**0.90**	0.60
surgery	0.50	0.12	**0.95**	0.07	0.02	0.02	0.02
transient	0.52	0.48	0.01	0.50	0.13	0.41	**0.97**
transport	0.53	0.31	0.01	0.65	0.87	0.69	**0.98**
ultrasound	0.43	0.19	0.17	**0.84**	**0.84**	0.81	0.80
variation	0.54	0.62	**0.80**	0.27	0.41	0.22	0.20
weight	0.51	0.40	0.45	0.51	0.43	**0.57**	0.55
white	0.49	**0.57**	0.54	**0.57**	**0.57**	0.54	**0.57**
Overall Accuracy	0.46	0.46	0.39	0.49	0.51	0.46	**0.53**

Table G.6: MetaMap Mapped Text Results using the Euclidean Distance

target word	Random	CUI50		CUI100		TERM 50		TERM 100	
	Baseline	o1	o2	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.32	0.22	0.23	0.20	0.44	0.19	**0.53**	0.29
blood pressure	0.38	0.41	0.34	0.36	0.29	0.23	**0.62**	0.21	0.39
cold	0.14	**0.41**	0.06	0.33	0.05	0.13	0.25	0.03	0.33
condition	0.54	0.54	0.14	0.49	**0.98**	0.88	0.02	0.95	0.03
culture	0.44	0.54	0.14	0.55	0.19	0.27	**0.75**	0.14	0.60
degree	0.49	0.23	**0.95**	0.18	**0.95**	0.29	0.60	0.20	0.00
depression	0.46	0.41	0.15	0.15	0.13	0.05	**0.93**	0.04	0.81
determination	0.44	0.22	**1.00**	0.44	0.67	0.38	**1.00**	0.59	0.71
discharge	0.40	**0.75**	0.08	0.59	0.16	0.53	0.67	0.49	0.24
energy	0.44	0.58	0.07	0.51	0.13	0.88	0.90	**0.93**	0.82
evaluation	0.52	**0.58**	0.48	0.57	0.56	0.46	0.50	0.53	0.55
extraction	0.43	0.62	0.06	**0.80**	0.08	0.61	0.06	0.59	0.06
failure	0.41	0.79	0.00	0.79	**0.86**	0.66	**0.86**	0.62	**0.86**
fat	0.51	0.41	**0.97**	0.49	0.84	0.29	0.74	0.11	0.37
fit	0.56	0.67	0.39	0.33	0.00	**0.78**	0.00	0.50	0.00
fluid	0.48	0.55	0.58	0.55	0.55	0.16	**1.00**	0.06	**1.00**
frequency	0.53	0.27	0.86	0.46	0.66	0.85	0.03	**0.90**	0.16
ganglion	0.52	**0.93**	0.14	**0.93**	0.27	**0.93**	0.54	**0.93**	0.12
glucose	0.54	0.64	0.42	**0.74**	0.34	0.40	0.09	0.60	0.10
growth	0.61	0.61	0.51	0.61	0.53	0.62	0.61	**0.63**	0.62
immunosuppression	0.48	0.52	0.44	0.41	0.53	0.63	0.51	**0.68**	0.59
implantation	0.49	0.54	**0.81**	0.42	0.76	0.39	0.23	0.33	0.30
inhibition	0.53	0.76	0.05	**0.83**	0.17	0.64	0.22	0.77	0.19
japanese	0.56	0.72	0.90	0.61	0.89	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	**0.90**	0.14	**0.90**	0.48	0.86	0.48	0.83	0.52
man	0.26	0.48	**0.72**	0.54	0.35	0.37	0.17	0.57	0.27
mole	0.39	0.18	0.08	0.12	0.29	0.68	**0.99**	0.62	**0.99**
mosaic	0.37	0.25	0.47	0.18	0.55	0.29	**0.56**	0.26	0.45
nutrition	0.42	0.39	**0.47**	**0.47**	0.33	0.33	0.44	0.38	0.43
pathology	0.45	0.53	0.13	0.48	0.23	**0.66**	0.57	**0.66**	0.63
pressure	0.28	0.46	**0.99**	0.34	0.93	0.55	0.86	0.35	0.93
radiation	0.52	0.53	0.60	0.56	**0.61**	0.43	0.60	0.37	0.60
reduction	0.36	0.45	**0.82**	0.45	**0.82**	0.73	0.18	0.73	0.18
repair	0.41	**0.65**	0.31	0.53	0.26	0.53	0.25	0.37	0.25
resistance	0.67	**1.00**	0.00	**1.00**	0.00	0.67	0.00	0.33	0.00
scale	0.32	0.46	0.00	0.45	0.22	**0.72**	0.02	0.68	0.11
secretion	0.53	0.20	0.01	0.28	0.01	0.21	0.01	**0.29**	0.01
sensitivity	0.31	0.59	0.00	0.47	0.04	0.45	0.16	0.45	**0.80**
sex	0.29	0.51	0.80	0.37	**0.81**	0.64	0.20	0.42	0.19
single	0.53	0.57	0.63	**0.68**	0.64	0.07	0.01	0.04	0.09
strains	0.49	**0.78**	0.35	0.74	0.48	0.31	0.01	0.45	0.01
support	**0.80**	0.40	0.20	0.30	0.20	0.30	**0.80**	0.20	0.40
surgery	0.50	0.18	0.50	0.24	0.39	0.19	**0.76**	0.24	0.46
transient	0.52	0.62	**0.68**	0.66	0.64	0.49	0.07	0.51	0.44
transport	0.53	0.20	**0.96**	0.32	0.05	0.73	0.03	0.83	0.09
ultrasound	0.43	0.77	0.17	0.82	0.20	0.53	**0.83**	0.52	**0.83**
variation	0.54	0.59	0.77	0.57	**0.79**	0.44	0.20	0.53	0.21
weight	0.51	0.43	**0.51**	0.49	0.43	**0.51**	**0.51**	0.49	0.47
white	0.49	0.39	0.44	0.47	0.52	0.43	0.58	0.43	**0.59**
Overall Accuracy	0.46	**0.52**	0.42	0.51	0.43	0.50	0.44	0.49	0.41

Table G.7: MetaMap Mapped Text Results using the Cosine Measure

target word	Random	CUI50		CUI100		TERM 50		TERM 100	
	Baseline	o1	o2	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.31	0.27	0.23	0.18	0.41	0.30	**0.51**	0.23
blood pressure	0.38	0.42	0.09	0.32	0.04	0.26	**0.53**	0.19	0.48
cold	0.14	**0.44**	0.13	0.43	0.08	0.16	0.36	0.03	0.29
condition	0.54	0.86	**0.97**	0.95	**0.97**	0.74	0.87	0.86	0.78
culture	0.44	0.51	0.38	**0.54**	0.43	0.26	0.41	0.14	0.33
degree	0.49	0.28	**0.45**	0.20	0.43	0.29	0.23	0.20	0.20
depression	0.46	0.41	**0.45**	0.15	0.18	0.06	0.09	0.11	0.02
determination	0.44	0.27	0.54	0.32	0.52	0.43	**0.58**	0.39	0.47
discharge	0.40	0.75	0.23	0.59	0.49	0.53	**0.88**	0.49	0.41
energy	0.44	0.61	0.77	0.68	0.46	0.88	**0.98**	0.91	**0.98**
evaluation	0.52	**0.59**	0.46	0.53	0.58	0.45	0.52	0.51	0.56
extraction	0.43	0.60	0.17	**0.66**	0.38	0.58	0.18	0.50	0.15
failure	0.41	**0.69**	0.17	**0.69**	0.17	0.66	0.34	0.66	0.31
fat	0.51	0.26	0.53	0.41	0.33	0.25	0.58	0.16	**0.60**
fit	0.56	0.78	0.67	0.56	0.06	**0.83**	0.44	0.50	0.22
fluid	0.48	0.60	0.69	0.66	0.75	0.34	**0.85**	0.17	0.81
frequency	0.53	0.04	0.00	0.02	0.00	0.74	0.80	0.70	**0.81**
ganglion	0.52	**0.93**	0.34	**0.93**	0.43	**0.93**	0.33	**0.93**	0.23
glucose	0.54	0.64	**0.74**	0.70	0.42	0.40	0.11	0.53	0.12
growth	0.61	0.61	0.58	0.61	0.58	0.62	**0.63**	**0.63**	**0.63**
immunosuppression	0.48	0.52	0.46	0.41	0.67	0.63	0.63	**0.70**	0.65
implantation	0.49	0.54	**0.69**	0.42	0.67	0.39	0.41	0.32	0.36
inhibition	0.53	0.76	0.53	**0.83**	0.57	0.62	0.75	0.73	0.47
japanese	0.56	0.81	0.62	0.84	0.51	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	**0.93**	0.10	**0.93**	0.14	0.90	0.07	**0.93**	0.07
man	0.26	0.42	**0.77**	0.46	0.35	0.28	0.15	0.47	0.08
mole	0.39	0.27	0.93	0.10	0.93	0.79	0.96	0.92	**0.98**
mosaic	0.37	0.35	0.53	0.37	**0.58**	0.38	0.53	0.42	0.49
nutrition	0.42	0.39	0.43	0.36	0.27	0.27	**0.45**	0.35	0.40
pathology	0.45	0.55	0.57	0.55	0.75	0.66	**0.86**	0.60	0.61
pressure	0.28	0.46	0.79	0.39	0.36	0.61	0.50	0.39	**0.83**
radiation	0.52	0.53	0.65	0.52	**0.69**	0.52	0.33	0.54	0.38
reduction	0.36	0.18	0.18	0.18	0.18	**0.73**	0.27	**0.73**	**0.73**
repair	0.41	0.65	**0.78**	0.53	0.41	0.56	0.60	0.54	0.28
resistance	0.67	**1.00**	**1.00**	**1.00**	**1.00**	0.67	**1.00**	0.33	**1.00**
scale	0.32	0.62	**0.91**	0.65	0.58	0.74	0.37	0.68	0.38
secretion	0.53	0.23	0.16	0.29	0.11	0.29	**0.37**	0.27	0.12
sensitivity	0.31	**0.69**	0.57	0.67	0.63	0.47	0.33	0.45	0.49
sex	0.29	0.51	0.48	0.41	0.55	**0.65**	0.51	0.58	0.35
single	0.53	**0.40**	0.09	0.37	0.09	0.07	0.05	0.08	0.09
strains	0.49	0.81	0.82	0.78	**0.91**	0.34	0.68	0.35	0.72
support	0.80	0.20	0.20	**0.40**	0.20	0.30	0.20	0.20	0.10
surgery	0.50	0.02	0.02	**0.03**	0.02	0.02	0.02	**0.03**	0.02
transient	0.52	0.59	0.27	**0.61**	0.26	0.40	0.18	0.40	0.12
transport	0.53	0.20	0.81	0.32	0.36	0.78	0.30	**0.85**	0.31
ultrasound	0.43	**0.76**	0.29	0.74	0.38	0.67	0.22	0.74	0.25
variation	0.54	**0.77**	0.72	0.75	0.75	0.32	0.28	0.43	0.33
weight	0.51	0.45	**0.62**	0.49	0.51	0.49	0.49	0.57	0.49
white	0.49	0.39	0.51	0.47	0.47	0.43	0.54	0.43	**0.59**
Overall Accuracy	0.46	**0.52**	0.49	0.51	0.44	0.50	0.47	0.49	0.43

Table G.8: MetaMap Mapped Text Results using the Dice Coefficient

target word	Random	CUI50		CUI100		TERM 50		TERM 100	
	Baseline	o1	o2	o1	o2	o1	o2	o1	o2
adjustment	0.27	0.30	0.35	0.26	0.55	0.53	0.60	0.58	**0.65**
blood pressure	0.38	**0.46**	0.35	**0.46**	0.14	0.16	0.42	0.07	0.45
cold	0.14	0.38	**0.48**	0.38	0.26	0.04	0.07	0.01	0.07
condition	0.54	0.95	**0.97**	**0.97**	**0.97**	0.79	0.86	0.90	0.71
culture	0.44	0.43	0.48	0.58	**0.86**	0.13	0.13	0.16	0.12
degree	0.49	0.22	0.08	0.14	0.08	0.22	**0.37**	0.12	0.11
depression	0.46	0.36	0.42	0.14	**0.79**	0.08	0.06	0.09	0.13
determination	0.44	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
discharge	0.40	**0.83**	0.61	0.69	0.81	0.59	0.64	0.52	0.15
energy	0.44	0.59	0.09	0.51	0.14	0.85	**0.99**	0.95	**0.99**
evaluation	0.52	**0.59**	0.50	0.47	0.49	0.48	0.51	0.47	0.48
extraction	0.43	0.27	**0.88**	0.45	0.86	0.30	0.08	0.38	0.33
failure	0.41	**0.86**	0.83	**0.86**	0.83	0.83	0.83	0.83	0.83
fat	0.51	**0.44**	0.21	0.27	0.05	0.18	0.12	0.07	0.12
fit	0.56	0.72	0.44	0.39	0.44	0.78	0.83	0.72	**0.94**
fluid	0.48	0.66	0.14	0.48	0.45	0.33	**0.84**	0.30	0.50
frequency	0.53	0.00	0.00	0.00	0.00	0.78	**0.89**	0.83	0.73
ganglion	0.52	0.93	**0.97**	0.93	0.93	0.93	**0.97**	0.93	0.92
glucose	0.54	0.67	0.42	0.67	0.60	0.48	0.52	**0.71**	0.56
growth	0.61	0.52	0.53	0.54	0.38	**0.63**	**0.63**	**0.63**	**0.63**
immunosuppression	0.48	0.44	0.55	0.40	0.47	0.60	0.48	**0.68**	0.48
implantation	0.49	**0.65**	0.62	0.41	0.52	0.53	0.50	0.45	0.37
inhibition	0.53	0.77	0.39	0.85	0.16	0.71	**0.99**	0.76	0.62
japanese	0.56	0.78	0.44	0.81	0.37	**0.94**	**0.94**	**0.94**	**0.94**
lead	0.21	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
man	0.26	0.35	0.45	**0.54**	0.23	0.27	0.29	0.42	0.26
mole	0.39	0.51	0.62	0.40	0.24	0.96	**0.99**	**0.99**	**0.99**
mosaic	0.37	0.25	0.52	0.26	0.49	0.46	**0.64**	0.51	0.60
nutrition	0.42	0.44	0.37	0.45	**0.47**	0.29	0.39	0.29	0.39
pathology	0.45	0.67	0.35	0.59	0.29	0.56	**0.81**	0.59	**0.81**
pressure	0.28	0.56	0.29	0.38	0.22	0.35	0.32	0.57	**0.90**
radiation	0.52	0.54	**0.61**	0.45	0.55	0.57	0.58	0.58	0.58
reduction	0.36	0.18	0.18	0.18	0.18	**0.73**	0.64	0.64	0.36
repair	0.41	0.62	0.63	0.40	0.41	**0.68**	0.34	0.49	0.31
resistance	0.67	**1.00**	**1.00**	**1.00**	**1.00**	0.67	**1.00**	0.33	**1.00**
scale	0.32	0.42	0.22	0.45	0.35	0.57	0.49	**0.66**	0.45
secretion	0.53	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
sensitivity	0.31	**0.59**	0.06	0.25	0.04	0.29	0.04	0.31	0.10
sex	0.29	0.50	0.30	0.50	0.41	**0.62**	0.28	**0.62**	0.19
single	0.53	0.50	0.64	0.54	**0.88**	0.07	0.01	0.08	0.01
strains	0.49	**0.43**	0.03	0.19	0.04	0.05	0.08	0.03	0.08
support	0.80	0.20	**0.60**	0.30	0.50	0.30	0.20	0.30	0.20
surgery	0.50	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**
transient	0.52	0.55	**0.69**	0.55	0.13	0.25	0.01	0.33	0.01
transport	0.53	0.19	0.57	0.30	0.12	0.88	**0.93**	0.91	0.88
ultrasound	0.43	0.76	0.74	0.74	0.68	0.78	0.75	**0.80**	0.73
variation	0.54	0.78	**0.80**	**0.80**	**0.80**	0.30	0.50	0.36	0.51
weight	0.51	0.66	0.47	0.58	0.45	0.53	0.62	**0.68**	0.45
white	0.49	0.52	**0.60**	0.50	0.58	0.43	0.46	0.46	0.44
Overall Accuracy	0.46	**0.51**	0.46	0.47	0.43	0.48	0.50	0.49	0.47