2002

TR-2002011: Corpus-Based Ambiguity Resolution of Biomedical Terms Using Knowledge Bases and Machine Learning

Hongfang Liu

Corpus-based Ambiguity Resolution of Biomedical Terms Using Knowledge Bases and Machine Learning

by

Hongfang Liu

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2002

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

_____	_____
Date	Chair of Examining Committee
_____	_____
Date	Executive Officer

Dr. Carol Friedman (Chair)

Dr. Virginia Teller

Dr. Ted Brown

Dr. Stephen Johnson

_____

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Corpus-based Ambiguity Resolution of Biomedical Terms Using Knowledge Bases and Machine Learning

by

Hongfang Liu

Advisor: Professor Carol Friedman

With the widespread use of natural language processing (NLP) techniques for information extraction and concept indexing in the biomedical domain, a method that efficiently and accurately assigns the correct sense of an ambiguous biomedical term in a given context is needed concurrently. The current status of resolving ambiguity in the biomedical domain is that handcrafted rules are used based on contextual material. The disadvantages of this approach are i) generating disambiguation rules manually is a time-consuming and tedious task, ii) maintenance of rule sets becomes increasingly difficult over time, and iii) handcrafted rules are often incomplete and perform poorly in new domains comprised of specialized vocabularies and different genres of text. We propose a two-phase method to build a classifier for an ambiguous biomedical term *W*. The first phase automatically creates a sense-tagged corpus for *W* using a biomedical terminology knowledge base, the UMLS, and free-text databases, and may include a semi-automatic process using clustering analysis and human supervision when we cannot automatically extract enough sense-tagged instances for *W*. The second phase automatically derives a classifier for *W* through supervised machine learning techniques using the derived sense-

tagged corpus as a training set. Experimental results show that generally the method can be used to construct WSD classifiers for abbreviations with a high precision without the need of human supervision. It can be used to construct WSD classifiers for general biomedical terms with a set of unrelated senses with a high precision when there are enough instances extracted for each sense. Clustering analysis can reduce human annotation cost when human supervision is needed.

Acknowledgements

This thesis was completed with the help of many people. The first and the most significant person that I would like to express my gratitude to is my advisor, Prof. Carol Friedman. I am indebted to her for giving me the encouragement, guidance and all the support I needed throughout the course of my Ph.D. study. From her broad technical knowledge combined with an optimistic and patient personality, I have learned the art of doing research, writing and presenting technical papers. It has indeed been a very precious experience to work with her. She influences me in many respects, especially in my attitude of doing research. She definitely sets a very good example for me to follow in my future research career. I could not have asked for a better advisor.

The other members of my thesis committee also deserve particular recognition. Prof. Virginia Teller offered an excellent course, Computational Linguistics, and introduced the field to me—thank you. Without that course, this thesis may never be possible. I would like to thank Prof. Ted Brown for providing me much freedom in the beginning of my Ph.D. study and for his belief in me. Prof. Stephen Johnson was invaluable for commenting on my writing and inspiring me through discussions with him and his research.

I thank all the people in the Department of Medical Informatics of Columbia University, especially, Prof. Soumitra Sengupta and Ms. Maria Zhang, for supplying computer resources, Prof. Ted Shortliffe and Ms. Rita Lenertz, for providing me an office space, Prof. Yves Lussier and Mr. Jianhua Li, for expert knowledge input. The research could not be done without their support. I would also like to thank Dr. Frank Hsu, for teaching

me how to be a researcher during my master study, and Prof. Stanley Habib, for admitting me, helping me during his time in Graduate Center, and his encouragement all these years.

Only the words from my heart can pay my debt of gratitude to my parents, Chaohuan Liu and Shiying Xia, for their love, understanding, and constant support. They kept answering the question asked by people, “why is Hongfang still studying instead of looking for a job”. Thanks to my son, Nan, for forgiving his mommy coming home late, keeping awake and welcoming mommy home with his bright smiling face every night. Last of all, thanks to my husband, Mao, for putting his own research aside and taking care of Nan when I was working late.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Motivation

1.1.1. Sense Ambiguity in the Biomedical Domain

With the widespread use of computers in the biomedical domain, a vast, rich range of biomedical data including coded data as well as free-text data has been stored in digital format. Computer applications can interpret coded data automatically while free-text data pose challenges to system developers. To enable access to free text in the biomedical domain, natural language processing (NLP) systems have been developed that facilitate information retrieval, information extraction, and text mining on free text [7;33;103]. However, all NLP systems require identification of terms (a term can be a single word or a multi-word phrase) in free text with entries in a lexical table [46;103]. Terms in free text can be ambiguous and may have multiple unrelated or related senses in the lexical table. For example, *capsule* can mean a unit for medication such as in "He was put on Dyazide one capsule daily over the past two days" or body region such as in "There may be faint lucency in the left internal capsule". The chemical term *potassium* can mean a laboratory test item in "Her potassium had been as low as 2.7 on July 27" or a drug item in "Her discharge medications are digoxin five days a week and potassium supplements 10 mEq each week day". It can also be an abbreviation that has multiple full forms or has the same spelling as a general English word, such as *HR,* which denotes hour or heart rate; and *SOB,* which denotes short of breath besides the general English word sob.

The need for resolving term ambiguity has been realized in NLP applications in the biomedical domain. Aronson [7] found that sense ambiguity resolution was important for

improving the performance of MetaMap, a free text to concept mapping program. An information extraction system, MedLEE, which was originally developed for radiology reports, encountered the sense ambiguity problem when broadened to a larger domain [31]. Nadkarni et al. [82] concluded that completely automated concept indexing in medical reports cannot be achieved without resolving sense ambiguities in free text. In an automatic knowledge discovery system, DAD-system, Weeber et al. [113] stated that in order to replicate Swanson's literature-based discovery of the involvement of magnesium deficiency in migraine [106], it was important to resolve the ambiguity of an ambiguous abbreviation *mg*, which denotes magnesium or milligram.

1.1.2. Lack of WSD Research in the Biomedical Domain

The task of resolving the ambiguity of ambiguous terms is called "word sense disambiguation" (WSD). A WSD system identifies the intended sense of a term in a context [86] from a set of candidates. Usually, a WSD system consists of a group of WSD classifiers, where each classifier determines the sense of a particular ambiguous term in a given context [45;86].

A WSD system that resolves sense ambiguities is essential for improving the precision of NLP applications in the biomedical domain. Several preliminary WSD methods for NLP applications in the domain were based on handcrafted rules. Rindflesch and Aronson [95] used a set of handcrafted rules based on semantic types of neighboring words to resolve ambiguity when mapping free text to UMLS concepts. The MedLEE system applies disambiguation rules based on local contextual information [31]. Johnson handcrafted rules to reduce ambiguity in a semantic lexicon for a particular domain [46]. It is expensive and difficult to write a comprehensive set of disambiguation rules.

Additionally, manual maintenance and further extension of rule sets become increasingly complex.

WSD systems have been built in the general English domain but are not suitable for NLP applications in the biomedical domain. An evident reason is the difference of "sense inventories". Words in the biomedical domain can take very restricted and specific meanings. For example, there are three senses of *discharge* in a biomedical terminology knowledge base, the UMLS [1], while there are nine senses for the noun *discharge* in an online general English lexicon WordNet [29]. In addition, for each specific domain, there exist a large number of terms that are exclusively used in that domain [120]. For example, abbreviations are widely used in medical reports for the reason that time pressure prevents medical specialists from describing clinical findings fully; and abbreviations are a convenient way to represent long medical words and phrases explicitly [8;71]. Another evident reason is the difference of contextual construction rules. Sentences are constructed in a more concise way in the medical domain than in some other domains: unnecessary words are usually omitted; questions are seldom found in medical reports; and verbs in discharge summaries usually appear as the third person singular [103].

1.2. Research Summary

In this dissertation, we propose a two-phase method to construct a WSD system, which consists of a set of classifiers (one for each ambiguous term), for NLP applications in the biomedical domain. Given an ambiguous term *W*, the first phase derives a sense-tagged corpus, STC(*W*), from a free-text collection based on a biomedical terminology knowledge base (the UMLS), machine learning techniques, and expert knowledge if

needed[1]. The second phase automatically constructs a WSD classifier for *W* by defining a supervised learning task that uses STC(*W*) as the training set.

Our solution is motivated by the following observations:

1. A concept-oriented machine-readable dictionary and a large size free-text collection exist along with NLP applications in the biomedical domain. Different electronic biomedical terminologies have been integrated into one concept-oriented knowledge base, the UMLS. Data repositories involving de-identified electronic medical reports offer a large size free-text collection for NLP applications in the clinical domain. The digitalization of journal articles and documents makes a large size literature collection possible.

2. Similar contexts imply similar senses. Similarity-based machine learning techniques have been applied to develop WSD systems and achieved good performance [49;85]. As a corollary of the observation, novel contexts imply novel senses or novel genres of the context. Corpus-based lexicographers have used this observation to discover new senses for a term [56].

3. A supervised machine learning approach using a large sense-tagged corpus is a viable way to build a robust, wide coverage, and highly accurate WSD system [84]. The observation is based on the success of using supervised machine learning approaches to automate relatively low-level language processing such as part-of speech tagging and segmenting text both in the general English domain and in the biomedical domain [9;13;20;92].

[1] A sense-tagged corpus for *W* is a collection of instances of *W* where the sense of W in these instances has been tagged, where an instance of *W* is a contextual unit that contains *W*. The content of a contextual unit depends on the genre of a collection and the exact application; it can be a sentence, two consecutive sentences, a paragraph, or a document, etc.

1.2.1. Derivation of a Sense-Tagged Corpus for *W*

Figure 1 illustrates the overall process of the derivation of a sense-tagged corpus for *W*, STC(*W*). There are four components in this phase. The first component, instance extractor, extracts two raw corpora: $RC_1(W)$ and $RC_2(W)$ from a free-text collection, where $RC_1(W)$ gathers instances that contain unambiguous synonyms of *W*, and $RC_2(W)$ contains instances that contain *W*. These are utilized by the second component, which derives a preliminary sense-tagged corpus, STC'(*W*). The third component is an optional component which automatically checks the quality of STC'(*W*) by applying clustering analysis. If STC'(*W*) is not found to have high quality, the fourth component (i.e., human annotation) is performed on the result of clustering.

The automatic derivation of a preliminary sense-tagged corpus for *W*, STC'(*W*), is based on two observations concerning WSD work that uses conceptual relations[2] in a machine-readable dictionary. One is that terms with similar senses tend to appear in similar contexts. The other observation is that the correct senses for words in a natural language expression will have closer sense relations than incorrect combinations of senses.

Based on the first observation, all instances that contain unambiguous synonyms (or closely related senses) of *W* are gathered, forming a collection $RC_1(W)$. A sense-tagged corpus $STC_1(W)$ is derived by replacing each occurrence of the corresponding unambiguous synonym with *W* and tagging the sense of *W* with the associated sense of that synonym for each instance in $RC_1(W)$. Therefore, instances in $STC_1(W)$ are actually

[2] A conceptual relation is a relation between two concepts that indicates these two concepts are related through some relations in the machine readable dictionary, such as the discharge substance sense of *discharge* is a child of the concept *substance*.

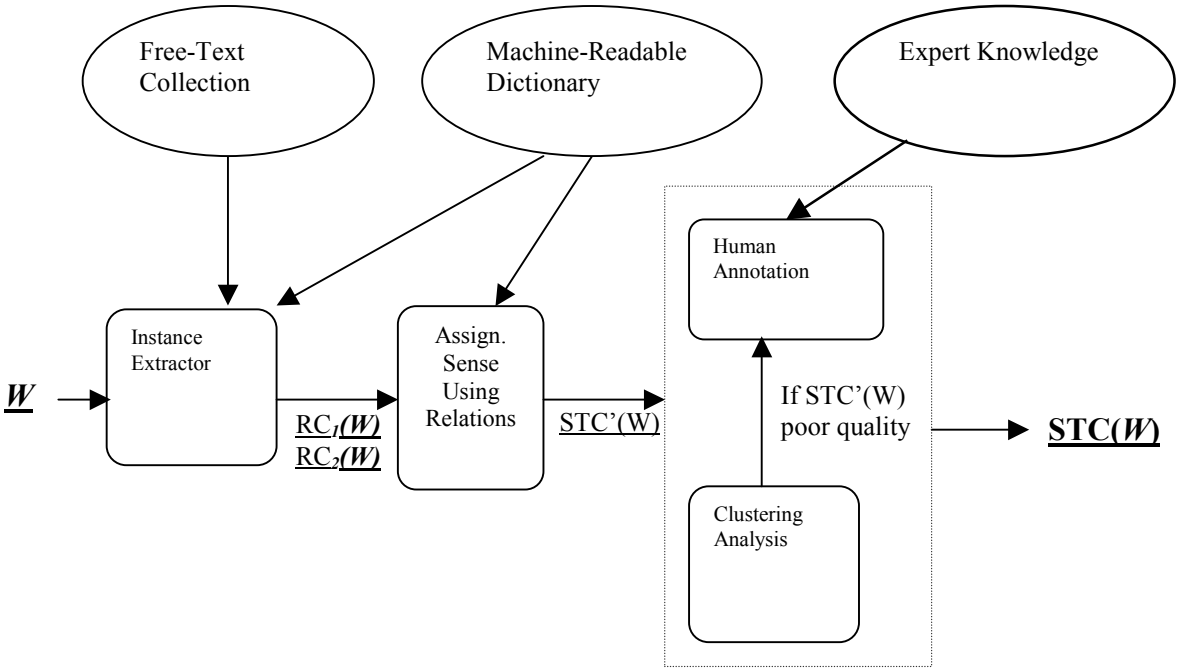


Figure 1. The process of construction of a sense-tagged corpus, STC(*W*), for a specified ambiguous term *W*.

artificial instances, i.e., they are derived from instances that contain synonyms of *W* instead of from instances that contain *W*.

Based on the second observation, we collect all instances that contain *W*, and denote the collection $RC_2(W)$. Terms that have conceptual relations with *W* are identified in each instance from $RC_2(W)$. A sense-tagged corpus $STC_2(W)$ is derived based on those identified terms. Some instances in $RC_2(W)$ may not have terms that are conceptually related to *W* and therefore cannot be sense-tagged. In addition, some instances in $STC_2(W)$ may be sense-tagged incorrectly because terms in an instance may be conceptually related to incorrect senses of *W*.

The preliminary sense-tagged corpus STC'(W) is obtained by combining $STC_1(W)$ and $STC_2(W)$. The quality of STC'(W) depends on i) how similar it is between W and a corresponding synonym with respect to the context, ii) the precision of $STC_2(W)$, and iii) the comprehensiveness of STC'(W) with respect to senses and genres of the context in $RC_2(W)$.

Two optional components, i.e., the clustering analysis component and the human annotation component, can be included to derive the final sense-tagged corpus STC(W). The purpose of the clustering analysis is two-fold: i) to check the quality of STC'(W), and ii) to reduce human annotation cost. The input to the clustering analysis component is the entire instance collection derived for W (i.e., STC'(W) and the portion of $RC_2(W)$ that could not be sense-tagged). Similar instances are grouped together based on similarity measures and clustering criteria. The clustering criteria are controlled by sense-tagged instances. The number of clusters in the final clustering is restrained by what is considered to be an affordable cost of human supervision. If all clusters in the final clustering possess some sense-tagged instances, then STC'(W) is considered to be comprehensive and becomes the final sense-tagged corpus that will be presented without human supervision to the second phase (i.e., the supervised machine learning phase). Otherwise, clusters that contain a large number of instances but have no sense-tagged instances need to be manually sense-tagged by experts. The sense-tagged corpus presented to the second phase will then contain instances in STC'(W) and the manually sense-tagged instances. A large sense-tagged corpus can be derived when assigning each raw instance the majority sense of its associated cluster.

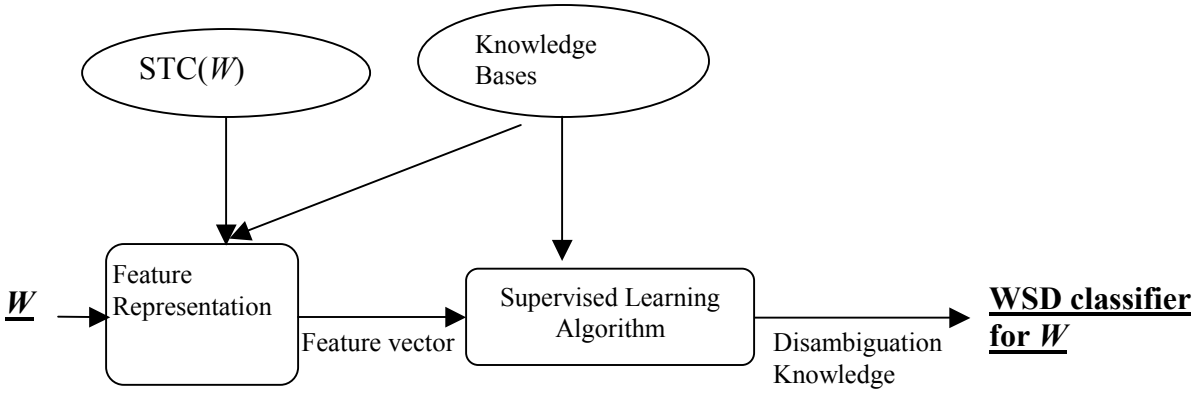


Figure 2. The processing phases for constructing a WSD classifier for *W*.

1.2.2. Construction of a WSD Classifier for *W*

Figure 2 demonstrates the construction process of a WSD classifier for *W* using the derived sense-tagged corpus, STC(*W*), as a training set. The input to the process is a sense-tagged corpus for *W* (STC(*W*)), and the output is a WSD classifier which can disambiguate *W*. The first component transfers each instance in STC(*W*) to a feature representation (usually a feature vector). It may utilize knowledge existing in knowledge bases for the transformation. The second component uses a supervised learning algorithm to learn disambiguation knowledge that forms a WSD classifier for *W*. It may also utilize knowledge existing in knowledge bases to choose an appropriate supervised learning algorithm and parameters of the algorithm.

Appropriate feature representations should capture features with high discrimination power, while the number of different features should be kept as small as possible in order to have classifiers with good generalization capabilities. Neighboring words and/or local collocations in a fixed window size are usually used to extract features. However, there is

no agreement on the best feature representation and the most appropriate window size. Several supervised learning methods have been adopted to built WSD classifiers: Naïve Bayes learning [11], neural networks [109], decision list [123], instance-based learning [45;86], inductive logic programming [80]. However, there is no agreement on the best choice of supervised learning algorithms.

1.2.3. Several Characteristics of WSD

There are several characteristics of WSD that must be addressed:

- No universal sense definition – there is no agreement about the set of senses for the same word in different lexicons. For example, *bank* has 9 noun senses in the Longman Dictionary of Contemporary English (LDOCE) [90] but 10 noun senses in an online general English lexicon WordNet [29], where two sets of sense definitions are not inclusive or exclusive as shown in Table 1 (e.g., Sense 9 in LDOCE matches Sense h in WordNet; there is no equivalence in WordNet for Sense 3 in LDOCE; and there is no counterpart in LDOCE for Sense j in WordNet).

We chose the UMLS as our machine-readable dictionary. The UMLS is the most comprehensive machine-readable dictionary in the biomedical domain and should be the most appropriate one for WSD in the biomedical domain. For example, two senses of *discharge* in the MedLEE lexicon (i.e., the discharge procedure sense as in "prior to discharge from hospital" and the discharge substance sense as in "bloody discharge") are both included in the UMLS while there is an additional sense for *discharge* in the UMLS: the discharge substance sample sense.

LDOCE	WordNet
1. Land along the side of a river, lake, etc 2. Earth which is heaped up in a field or garden 3. A mass of snow, clouds, mud etc 4. A slope made at bends in a road or race-track 5. A high underwater bank of sand 6. A row 7. A place in which money is kept 8. A place where something is held ready for use 9. The funds held by a gambling house	a. A financial institution that accepts deposits and channels the money into lending activities b. Sloping land c. A supply or stock held in reserve for future use d. A building in which commercial banking is transacted e. An arrangement of similar objects in a row or in tiers f. A container (usually with a slot in the top) for keeping money at home g. A long ridge or pile h. The funds held by a gambling house i. A slope in the turn of a road or track j. A flight maneuver

Table 1. The noun senses for *bank* in two different machine-readable dictionaries: LDOCE and WordNet.

- Low inter/intra agreement among subjects – evidence to date suggests that people often disagree on the sense to be assigned to an instance of a word [30;47]. Veronis [111] showed an average agreement measure using the Kappa statistics [15;22] was below 50%, which indicated a large amount of disagreement among subjects. Jorgensen [47] found the agreement level on the appropriate sense for an instance to be just 68%. Weeber and colleagues [114] were aware that 12 of 50 ambiguous terms

used in their WSD project were problematic; subjects often disagreed with each other on the sense assignment of those terms.

The cause of the low inter/intra agreement is the fine-grained sense-granularity of the corresponding sense inventory. As pointed out by Krovetz[59], Sanderson[98], and Chen & Chang[17], senses used in machine-readable dictionaries are usually too fine-grained for WSD systems and for NLP applications in the general English domain. The maintenance of the UMLS contains a knowledge engineering process with the input of expert knowledge. During the process, similar UMLS concepts are merged for the purpose of accurate information retrieval. For example, the 2002 version of the UMLS includes 776,940 concepts and 2.10 million concept names. Compared with the 2001 version, there were 20,419 fewer concepts, while there were 137,056 more strings. The UMLS keeps merging concepts together and becomes more and more appropriate for WSD and NLP. For example, in the 1999 version of the UMLS, *radiation* had three different senses while these three senses were merged to one sense in the 2001 version.

- Novelty—new words, new senses and new usages of existing words appear constantly [54]. With the explosive knowledge discovery in the biomedical domain, new terms are invented continuously; and new abbreviations and new senses of abbreviations (i.e. new full forms) come out daily. Observed by Cheung [18], the abbreviations used in clinical trials of cardiology alone increased from 200 in 1992 to 2,300 in 1998. In addition, the novelty also appears for NLP applications when the applications transfer (or broaden) to a new (or larger) domain. For example, in conjunction with the broadening process of MedLEE [31], the size of the MedLEE

lexicon has increased from an initial amount of 4,500 entries to a total of 15,307 entries.

Novelty in WSD presents limitations for handcrafted WSD rules since handcrafted rules are not easy to maintain and update. However, the sense inventory of our system, the UMLS, is updated annually. New words and new senses of existing words are likely to be included in future versions of the UMLS. New usages of existing words would most likely be noticed when using the most current version of free-text collections to derive sense-tagged corpora. The automatic derivation of sense-tagged corpora and the automatic construction of WSD classifiers in our method facilitate maintaining and updating our WSD system.

1.2.4. Overview of Experiments

We used three sets of ambiguous terms for the experiments.

- A - contains 35 frequently occurring ambiguous abbreviations in the medical reports;
- B - contains 38 general ambiguous terms used in the WSD project of National Library of Medicine (NLM), where the gold standard set was determined manually by Weeber and his colleagues [114];
- C - contains 4 ambiguous terms, i.e., *cold*, *discharge*, *lead*, and *dressing*, in the clinical domain. The gold standard set has been derived manually using human experts.

The experiments were designed for answering the following broad questions:

1. Given an ambiguous term *W* and a sense-tagged corpus for *W*, how to construct a supervised WSD classifier for *W*? As we know, there is no agreement on the preferred feature representation, the suitable window used to extract features, and the best supervised learning algorithm for WSD;

2. What kinds of terms can use our method to automatically derive WSD classifiers with a reliable precision, i.e., without two optional components (clustering analysis and expert annotation) in the first phase of our method? And what kinds of terms require expert annotation?

3. What is the conceptual coverage of the UMLS for biomedical terms? And is it feasible to automatically understand abbreviations in MEDLINE?

For the first question, we conducted a comparison study of supervised WSD classifiers using sets A and B (see Section 1.2.4.2 for an overview), where the gold standard sets were automatically derived from MEDLINE (see Section 1.2.4.1 for an overview). In addition, we compared the noise tolerance of different supervised learning algorithms (see Section 1.2.4.3 for an overview).

For the second question, we proposed several hypotheses.

Hypothesis 1. Our method can be used to automatically derive WSD classifiers for abbreviations in MEDLINE with a set of known full forms (also termed as expansions, or definitions). Note that automatic derivation here means that the method is used without the inclusion of two optional components: clustering analysis and expert annotation.

Hypothesis 2. WSD classifiers for abbreviations, which are trained on sense-tagged instances derived from MEDLINE, can also be used to disambiguate instances in the clinical domain.

Hypothesis 3. Our automatic extraction of sense-tagged instances can also be applied to derive sense-tagged instances for a majority of ambiguous UMLS biomedical terms.

Hypothesis 4. The derived WSD classifiers achieve a high precision for ambiguous UMLS biomedical terms without closely related senses provided there are enough instances.

Hypothesis 5. Clustering analysis can reduce human annotation cost dramatically.

Hypotheses 1 and 2 were proposed for abbreviations, and an overview of the associated experiments is presented in Section 1.2.4.4. Hypotheses 3, 4, and 5 were proposed for ambiguous general biomedical terms, and an overview of the corresponding experiments is given in Section 1.2.4.5.

The third question was assessed through several studies including a study of conceptual coverage of the UMLS and a study of MEDLINE abbreviations. Overviews are provided in Sections 1.2.4.6 and 1.2.4.7, respectively.

1.2.4.1. Automatic Derivation of Gold Standard Sets for Abbreviations

The evaluation of WSD classifiers requires gold standard sets that are very expensive to derive manually. However, abbreviations are usually defined in the literature, where senses of abbreviations are the same as the corresponding full forms. A gold standard set was automatically derived for each abbreviation in Set A from MEDLINE by omitting

the full form of an abbreviation in an instance and assigning the associated sense of the full form to the corresponding abbreviation. In addition, hundreds of gold standard instances were also derived from a collection of medical reports for abbreviations in Set A.

1.2.4.2. Comparison Study of Supervised WSD Classifiers

We did a thorough comparison study of supervised WSD classifiers with four variables: type of ambiguous terms, feature representation, supervised learning algorithm, and window size.

Two types of terms were used in the study: abbreviations in Set A and general ambiguous biomedical terms in Set B. Gold standard sets for abbreviations in Set A were automatically derived from MEDLINE as described in the previous section, and gold standard sets for terms in Set B were determined by Weeber and his colleague [114] for the WSD project of the National Library of Medicine. The feature representation variable had six different options: a) words with oriented distance within the window, b) words with orientation within the window, c) words within the window, d) three collocations, oriented words within a window of size 2, e) features in “c” and “d”, and f) features in “d” and all other words in the corresponding instance. Five different supervised learning algorithms were used including three existing algorithms (i.e. Naïve Bayes learning, traditional implementation of decision list learning, instance-based learning) and two new algorithms (i.e., our implementation of decision list learning and our mixed supervised learning). For abbreviations in Set A, we used three different window sizes: 3, 5, and 10. For general biomedical ambiguous terms in Set B, we used five different window sizes: 2, 3, 4, 5, and 10. Note that we could test every possible window size, we selected these

window sizes in order to see the preference of window sizes. The different selection of window sizes between Set A and Set B was purely determined by the time needed for the experiments (there were much more instances in Set A than in Set B, so we chose 3 different window sizes for Set A instead of 5 which were used for Set B).

We found that instance-based learning was very time-consuming for WSD when there were a large number of sense-tagged instances. We aborted all instance-based classifiers. We discovered that all supervised WSD classifiers had a reliable performance when there were hundreds of sense-tagged instances for each sense. Our mixed supervised learning was better than Naïve Bayes learning; and our implementation of decision list learning was better than traditional decision list learning. We also found that feature representations including collocations (i.e., an ordered set of words in context around the corresponding ambiguous term) and neighboring words (i.e., a set of words in context around the corresponding ambiguous term) were appropriate representations for the context. For terms with domain-specific senses such as abbreviations, a large window size, such as the whole instance, was promising. For general biomedical ambiguous terms, where the ambiguity might be caused by related senses, a small window size of 2 to 5 had a better performance.

1.2.4.3. Noise Tolerance of Different Supervised Learning Algorithms

We used abbreviations in Set A with the same gold standard sets as in the previous experiment to compare noise tolerance of four supervised learning algorithms (i.e., Naïve Bayes learning, traditional implementation of decision list learning, our implementation of decision list learning, and our mixed supervised learning). The gold standard set was divided into a training set and a test set with the ratio 9:1. Nine different levels of noise

(i.e., some instances in the training set were assigned the wrong senses) were introduced to the training set. Measures were averaged over 5 random runs.

The tolerance of noise was different among different supervised learning algorithms. Naïve Bayes learning could not tolerate noise. The precision of Naïve Bayes classifiers was very low when noise was present in the training set for abbreviations with a skewed sense distribution or with rare senses[3]. Our implementation of decision list learning had a lower degree of noise tolerance compared to traditional decision list learning. Our mixed supervised learning had the best performance for abbreviations with a balanced distribution for majority senses, while traditional decision list learning was robust and had the best performance for abbreviations with a skewed sense distribution.

1.2.4.4. Construction of WSD Classifiers for Abbreviations

Abbreviations appear frequently in the biomedical domain and they are frequently ambiguous. The correct interpretation of abbreviations is critical for NLP applications in the biomedical domain since most abbreviations hold domain-specific senses. A reliable WSD system in the biomedical domain should disambiguate abbreviations with a high precision. The proposed hypotheses 1 and 2 were regarding the construction of WSD classifiers for abbreviations.

We answered Hypothesis 1 through three experiments:

I. We built WSD classifiers for abbreviations in Set A using sense-tagged corpora that were derived from MEDLINE using unambiguous synonyms (STC_1). The

[3] A sense distribution is skewed if the majority sense has an occurrence of over 90% of the total, otherwise, the sense distribution is balanced; a rare sense is a sense with an occurrence of less than 20 and less than

performance of those WSD classifiers was evaluated using the automatic derived gold standard sets.

II. We evaluated the quality of sense-tagged corpora, which were derived from MEDLINE using conceptual relatives in the context.

III. We addressed the performance of WSD classifiers that were constructed automatically using knowledge acquired from previous experiments.

We discovered that for abbreviations with unrelated senses where the relatedness information was from the corresponding sense inventory, the constructed WSD classifiers had a high precision (around 97%). For abbreviations with a relatively balanced distribution for majority senses, our mixed supervised learning achieved the best performance; and for abbreviations with a skewed sense distribution, decision list learning implementations were the best.

As we know, abbreviations are frequently used in medical reports without definitions. The disambiguation of abbreviations is important for NLP applications in the clinical domain. From a previous study [69], we found that we could not derive enough instances for abbreviations from medical reports using synonyms. Furthermore, compared to MEDLINE abstracts, medical reports usually contain more than one topical concept. Our method that uses conceptual related terms in the context to derive sense-tagged corpus may not work well in the clinical domain. For example, the diagnosis section of a discharge summary may contain information about all diseases of the corresponding patient, while one MEDLINE abstract usually contains information about only one

---

1% of the total occurrence or with an occurrence of less than 0.5% of the total occurrence; otherwise, the

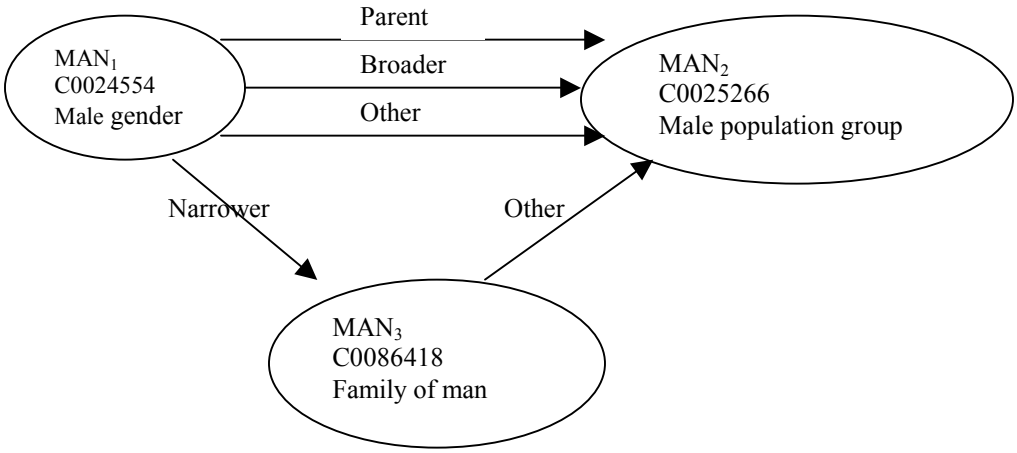


Figure 3. Three UMLS concepts denoted by the term *man* are closely related. The oval components are senses. The lines between them are relations defined in the UMLS together with the corresponding relation types.

disease and its relevant information. Hypothesis 2 stated that WSD classifiers constructed using instances from MEDLINE could be used to resolve ambiguity in the clinical domain.

The hypothesis was assessed by applying WSD classifiers for abbreviations in Set A, which were constructed using instances in MEDLINE, to disambiguate gold standard instances derived from medical reports. Results demonstrated a precision of 98%.

1.2.4.5. Construction of WSD Classifiers for General Biomedical Terms

For ambiguous biomedical terms that correspond to general English words, such as those in Set B, two optional components are unavoidable for the following reasons. First, not all senses of these terms are biomedical concepts, which implies WSD is not well defined when using the UMLS as the sense inventory. For example, the verb sense of *lead* as in

---

sense is a majority sense.

the sentence "Failure to recognize internal structures of populations may lead to considerable bias in predicting effective size" is not a biomedical concept, for which we cannot derive sense-tagged instances using the UMLS. Secondly, the sense granularity of the UMLS for some of these terms is too fine-grained. For example, three UMLS concepts (i.e., Male gender, Male population group, and Family of man) for the term *man* are closely related as shown in Figure 3, where each concept is represented using ovals and links between two ovals indicate relations defined in the UMLS.

We proposed three hypotheses (i.e., Hypotheses 3, 4, and 5. See the beginning of Section 1.2.4) for ambiguous general biomedical terms.

Hypothesis 3, which states that our automatic extraction of sense-tagged instances can be applied to derive sense-tagged instances for a majority of ambiguous UMLS biomedical terms, was tackled by measuring the number of ambiguous UMLS biomedical terms (i.e., strings listed in the UMLS ambiguous string table, which consists of 4,457 strings) with at least one concept that did not have conceptually related terms. If a concept has conceptually related terms, we can derive sense-tagged instances for it based on these related terms. The result showed that there were only 4 (out of 4,457) ambiguous terms with one concept that could not use our method to derive sense-tagged instances.

Two experiments were designed to address Hypotheses 4 and 5.

In the first experiment, we derived sense-tagged instances for each word in Set B. We applied clustering analysis to check the quality of the derived sense-tagged instances for words with unrelated senses and the corresponding STC_2 having a high precision.

In order to exclude biases caused by cases where there were not enough gold standard instances, we only built WSD classifiers for words where the gold standard set had a balanced sense distribution and the size of the gold standard set was over a large enough threshold (i.e., 15 here). The result demonstrated that the derived sense-tagged corpora for words that were considered were comprehensive. There were only two words, *white* and *implantation*, satisfied the criteria for building WSD classifiers. The precision of the constructed WSD classifier for *white* was 85.6% compared to a precision of 54.4% when assigning the majority sense to each instance. The precision of the constructed WSD classifier for *implantation* was 93.9% compared to 81% when assigning the majority sense to each instance.

In the second experiment, we acquired sentences for each word in Set C from a collection of medical reports. We then applied clustering analysis on sentences and derived a set of clusters. Human experts were asked to sense-annotate an instance randomly chosen from each cluster. The sense of that instance was then assigned to each instance in that cluster and a sense-tagged corpus was built for each word. A WSD classifier was constructed and tested on 50 gold standard instances. The results illustrated that the WSD classifiers achieved a precision of 98% or higher for all words except the word *cold* where the constructed WSD classifier had a precision of 86% and *cold* had five majority senses.

1.2.4.6. The Conceptual Coverage of the UMLS

Our automatic derivation of sense-tagged corpora is based on the UMLS. As we saw at the beginning of this section, there is no universal sense definition for the same term. The UMLS combines different electronic biomedical terminologies, to what extent. We wanted to evaluate to what extent the UMLS could serve as a sense inventory for a WSD

system in the biomedical domain. Three conceptual coverage studies were performed including a study of the UMLS abbreviation coverage of abbreviations in medical reports, a study of the UMLS coverage of the MedLEE lexicon, and the UMLS coverage of abbreviations in MEDLINE abstracts as included in the next study (See Section 1.2.4.7). Results demonstrated that for abbreviations, the conceptual coverage of the UMLS was related to the frequency. The UMLS covered around 80% of frequent abbreviations either from medical reports or from MEDLINE when the frequency was over 50. There were 54.7% of MedLEE lexicon entries that were automatically mapped to the UMLS together with the correct associated semantic categories.

1.2.4.7. The Study of MEDLINE Abbreviations

In Section 1.2.4.4, we have shown that our method can be used to construct WSD classifiers for abbreviations with a high precision. However, since not every full form of an abbreviation is included in the UMLS, we preformed a feasibility study of automatic understanding of abbreviations appearing in MEDLINE.

There are several steps for the automatic understanding of abbreviations. First, a method to associate an abbreviation to its corresponding full form in the context is needed, with an assumption that the authors define abbreviations when they are first introduced in a specific domain for the less well-known abbreviations. Secondly, well-known abbreviations are not always defined in documents. In order to understand these, an abbreviation database that lists abbreviations together with their senses needs to be built and updated periodically. However, manually constructing a database is time-consuming. In addition, manual maintenance and further extension are increasingly complex. But constructing an abbreviation database automatically by matching abbreviations with their

full forms in documents requires a method to group textual variants together and a method to link them to the proper sense. Finally, abbreviations are highly ambiguous. The number of characters that form an abbreviation is limited, and abbreviations are usually short. With the rapid growth of the use of abbreviations, one abbreviation may represent dozens of senses. A method to resolve the sense ambiguity is needed.

The feasibility study consists of answers for the following questions using three-letter abbreviations in MEDLINE abstracts: can we build an abbreviation knowledge base from MEDLINE abstracts? If yes, what is the UMLS concept coverage, what is the average number of textual variants for each sense, how ambiguous are the abbreviations, and what is the role of the frequency of the senses?

The results demonstrated that automatic understanding of abbreviations was feasible for frequently occurring abbreviations. After ignoring senses with less than 100 occurrences, over 80% of the senses matched the UMLS; 22.0% of the abbreviations were ambiguous, with an average of 2.36 senses for ambiguous ones, which could be resolved based on our previous experiments.

1.3. Research Contributions

The contributions include:

- This is the first systematic WSD work in the biomedical domain. Researchers in the computational linguistics field debate the soundness of treating WSD as a classification task as part of speech tagging, and the feasibility of building a universal WSD system[55;116].

- Large-scale evaluations of WSD systems are typically impeded by the lack of a gold standard set[56;57]. We provide a method for automatic evaluation of our WSD system using abbreviations.

- We provided a thorough comparison study of different supervised WSD classifiers with four variables: type of ambiguous terms, feature representation, supervised learning algorithm, and window size. We also compared the noise tolerance of different supervised learning algorithms.

- Our implementation of decision list learning, which separates features that occur with only one sense from other features, has a better performance than traditional implementations of decision list learning, which do not distinguish these two, when there is no noise in the training set.

- Traditional WSD implementations of Naïve Bayes learning do not distinguish rare senses from majority senses in the training set. We divided these two and proposed a mixed supervised learning algorithm that combines a Naïve Bayes classifier with an instance-based classifier using a local similarity measure (i.e., the computation of the similarity between two instances is only based on features of these two instances).

- This is the first large-scale WSD work that combines sense-tagged corpora derived using machine-readable dictionaries with supervised machine learning techniques. Previous WSD work isolates these two[4;77;105].

- We discovered that the best choice of window size depends on certain characteristics of the ambiguous terms. Domain-specific ambiguous terms require a large window

such as the whole instance, while general terms require a window of size 2 to 5. The best choice for a supervised learning algorithm depends on the sense distribution in the corresponding sense-tagged corpus. For terms with a sense-tagged corpus that is balanced among majority senses, our mixed supervised learning achieves the best performance; for a skewed sense-tagged corpus, traditional decision list learning achieves the best performance when there is noise in the training set; otherwise, our implementation of decision list learning achieves the best performance.

- We developed a clustering algorithm that can handle a large number of instances with a large number of features without the requirement of a pre-determined fixed number of clusters. Most existing clustering algorithms are optimized and suffer from either a speed or space problem [108]. We sacrifice a little bit of the clustering quality to solve these problems.

1.4. Outline

The remainder of the dissertation is organized as follows. Chapter 2 gives background information and previous work about WSD. General background information about resources used and NLP systems involved is provided in Chapter 3. Automatic derivation of sense-tagged corpora using relations and the clustering algorithm to check the quality of the automatically derived sense-tagged corpora and to reduce the human annotation cost are described in Chapter 4. Methods to automatically derive the gold standard set for Set A and detail information about three evaluation sets are presented in Chapter 5. A set of experiments (including the comparison study of supervised WSD classifiers, the noise tolerance study, and construction of WSD classifiers for abbreviations as well as for general biomedical terms) to evaluate the proposed method is presented in Chapter 6.

Chapter 7 discusses the conceptual coverage of the UMLS and the feasibility study of automatic understanding of MEDLINE abbreviations. Finally, Chapter 8 discusses implementation issues and concludes the dissertation.

Chapter 2. Background and Previous WSD Work

2.1. Previous WSD Research in the General English domain

The overview of previous WSD research presented here is divided into several topics: disambiguation knowledge sources, previous WSD work based on machine-readable dictionaries, early WSD work, feature representation, supervised WSD methods, unsupervised WSD methods, implementation issues, evaluation issues and several other issues.

2.1.1. Disambiguation Knowledge Sources

In the computational linguistics field, the disambiguation knowledge for WSD can be acquired automatically through two different sources [48]:

- Knowledge-bases, usually a machine readable dictionary (MRD), such as WordNet[29], Longman Dictionary of Contemporary English[90], Roget's thesaurus[16], etc.;

- Corpora including manually assembled sense-tagged corpora (e.g. the Semcor corpus[29], the DSO corpus[85]) or raw corpora (e.g. the Brown Corpus[62] and the BNC Corpus[4]); a WSD method using raw corpora only is not strictly a WSD method (since senses assigned to each instance are not well defined), and is usually referred as sense discrimination[100].

Machine-readable dictionaries here include online ordinary dictionaries, thesauri, and semantic lexicons. Ordinary dictionaries such as LDOCE[90] and CED[38] consist of an

[4] See http://www.hcu.ox.ac.uk/BNC/

alphabetical list of words with sense definitions in terms of other words, subject codes, and often several example usages, etc. Thesauri such as Roget's Thesaurus[16] provide semantic categories and information about relations among words in a highly systematic structure, where relations usually include synonymy relations (e.g. *discharge* via *emission*) and hypernymy relations (i.e., IS-A relation, e.g., *crane* is a kind of *machine*). Semantic lexicons in the general English domain refer to WordNet[29], which includes sets of synonymous words, called as synsets. Each synset is defined using a sense definition in terms of other words, comments and examples. In addition, synsets are linked into conceptual networks through different kinds of relations including synonymy, antonymy (i.e. opposite relation, e.g. *big* v.s. *small*), hypernymy, and meronymy (i.e., PART-OF relation, e.g., the *root* is a part of a *tree*), etc.

There are several corpora that are frequently used in corpus-based NLP research. The most famous corpus is the Brown Corpus that consists of a 1-million-word collection of instances from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc.) assembled at Brown University in 1963-1964[62]. The Wall Street Journal (WSJ) corpus is one of the corpora collected by the Association for Computational Linguistics's Data Collection Initiative (ACL/DCI)[5]. It consists of news stories from three-year WSJ archive (1987-1989) with about 30 million words of text. The British National Corpus (BNC) was carried out and is managed by an industrial/academic consortium lead by Oxford University Press. The corpus consists of over 100 million words, where 90% is written text (extracted from newspapers, journals, academic books and popular fictions, etc.) and 10% is spoken text (transcribed from

[5] See http://www.ldc.upenn.edu/Catalog/LDC2000T43.html

informal conversations, formal business or government meetings, radio shows and phone-ins, etc).

There are several sense-tagged corpora that have been used extensively in the literature. The most widely used sense-tagged corpus is SEMCOR. It comprises 250,000 words (taken from the Brown Corpus and a novel, The Red Badge of Courage) in which all content words have been sense-tagged using WordNet as sense inventory. DSO is another large size sense-tagged corpus where all occurrences of 191 "most frequently occurring and most ambiguous" nouns and verbs from the Brown corpus and a portion of the WSJ corpus were manually tagged using WordNet senses. Several small-scale sense-tagged corpora are also available through different research groups. These include a corpus of 2,094 examples with 6 senses of the noun *line*[65], a corpus of 2,369 sentences with 6 senses of the noun *interest*[11], and a corpus of 6,197 samples with 25 very high-frequency verbs[12].

2.1.2. Previous WSD Work based on Machine-Readable Dictionaries

An early WSD implementation of machine-readable dictionaries is due to Lesk[66], where the Oxford Advanced Learners Dictionary (OALD) was used. He applied a simple approach by counting the overlap between words used in the definitions of the senses. For example, Lesk's program correctly identifies the sense of *pine* as $pine_1$ and the sense of *cone* as $cone_3$ in the phrase "pine cone" given the following definitions[6]:

- $pine_1$: kind of evergreen tree with needle-shaped leaves
- $pine_2$: waste away through sorrow or illness

[6] "*evergreen*" and "*tree*" are two overlap words between $pine_1$ and $cone_3$

- $cone_1$: solid body which narrows to a point
- $cone_2$: something of this shape whether solid or hollow
- $cone_3$: fruit of certain evergreen trees

Lesk reports accuracies of 50-70% in disambiguation of the words in a small collection of text.

Wilks et al. [117] attempted to improve the knowledge associated with each sense definition by calculating the frequency of co-occurrence for the words in definition texts, from which they derived several measures for the degree of relatedness among words. This metric was then used to compute the similarity of each sense definition and a given context. Using a handcrafted gold standard set consisting of 197 occurrences of *bank*, the method achieved a precision of 45% using fine-grained senses, and 90% using coarse-grained senses. Veronis and Ide [44;112] extended the method by creating a neural network from definition texts in the Collins English Dictionary (CED), in which each word is linked to its senses, which are themselves linked to the words in their definitions, which are in turn linked to their senses, etc. An experiment on 138 instances for 23 ambiguous words with 6 instances for each, the method correctly disambiguated 71.7% of the occurrences using fine-grained senses, and 85% of the occurrences using coarse-grained senses. However, sense definitions are predefined and limited; so many WSD methods combine sense definitions with corpora (see Section 2.1.4.4).

The earliest example of the use of semantic categories is the work of Masterman [74] on machine translation. She associated each Latin word stem with its English equivalence through categories in Roget thesaurus [16]. Subject codes (which are roughly equivalent

to semantic categories) in sense definitions of many dictionaries have also been used together with the definition text. For example, the entry for *bank* in LDOCE includes the subject code EC (i.e. Economics) for the financial senses of *bank*. Cowie et al. [23] combined Lesk's method with subject codes and reported results of 47% for fine-grained senses and 72% for coarse-grained senses. However, subject codes are usually problematic and incomplete. Krovetz [59] used domain labels, i.e., domain information indicated within parentheses (e.g. "penalty--(in sports) a disadvantage given to a player or team for breaking a rule") to measure the quality of subject codes. He found that among 620 instances that contained domain labels and subject codes, 2% of occurrences were assigned wrong subject codes, and 4% of occurrences missed some subject codes.

Conceptual relations defined in machine-readable dictionaries are also used for WSD under the following observation: the correct senses for the words in a natural language expression will have closer sense relations (in a conceptual network) than incorrect combinations of senses [3;4;105]. For instance, in "Spring is my favorite season", the springtime sense of *spring* has a IS-A relation with the season of the year, while any other combination of senses (e.g. *spring* as a fountain and *season* as sports season) have weaker relationships. The corresponding WSD method consists of looking up terms that have relations with *W* in the context of *W*. The method takes a number of terms via a relation and a formula to measure the relatedness of those terms with each of the senses of *W* in a machine-readable dictionary, and then uses them to determine the sense of *W* in the context. In the general English domain, researchers usually choose WordNet as the concept-oriented dictionary. Sussna [105] used several relation types (such as hyponymy, synonymy etc) in WordNet, and chose a measure that takes account of the shortest path,

the number of edges with the same type leaving a node, the depth of a given edge in the overall tree, and a weight assignment for each relation type. Sussna evaluated his method on five documents from TIME magazine by comparing it to human experts using the same evidence and achieved a precision of 52.3%. Agirre and Rigau [4] proposed a method that used conceptual relatives via the relation IS-A, and chose a measure which is sensitive to the following parameters: the length of the shortest path that connects the concepts involved, the depth in the hierarchy, and the density of concepts in the hierarchy. Agirre and Rigau [3] evaluated their method on the noun portion of a document that contained 2,079 words. The overall performance was measured in terms of precision and recall with 66.4% for precision and 58.8% for recall.

2.1.3. Early WSD Work based on Corpora

An early example of using corpora in WSD is the research of Weiss [115] who constructed a set of rules manually based on the statistical information gathered from dozens of manually sense-tagged sentences. Words are disambiguated via two kinds of rules: template rules and general context rules. Template rules were learned using words within a window of size 2; and general context rules were learned using words within a window of size 5. For example, the following are two rules for the word *type*: if "*of*" appears immediately after *type* in a context, then *type* in the context means a particular kind of thing; if *pica* or *print* appears within a window of size 5 in a context, then *type* in the context is given a printing interpretation. The method was tested on five ambiguous words with a training set of 20 sentences each, and a test set of 30 sentences. The precision was about 90%. Later, Kelly and Stone [51] extracted concordances for 1,800 ambiguous words from a corpus of a half-million words. The concordances served as a

basis for the manual creation of a set of rules for the disambiguation of each of the 1,800 words. The rules consist of context rules that are similar to general context rules used by Weiss. In addition, there are some grammar rules that examine syntactic information. The rules are grouped into sets so that only certain rules are applied in certain situations. Their rules were tested and achieved a precision of 92% for coarse-grained sense distinctions.

2.1.4. Machine Learning WSD Methods

In the 1980s, with the development of computer and information sciences, many large-scale electronic corpora become available. More recent WSD approaches have shifted to an empirical paradigm where classifiers are constructed through machine learning using a large corpus of training data instead of manually handcrafting classifiers. The disambiguation knowledge can be acquired by applying machine learning algorithms on manually sense-tagged corpora or raw corpora that are combined with machine-readable dictionaries.

2.1.4.1. Background of Machine Learning

Machine learning is an automatic process of the construction of certain classifiers from a large collection of instances, which can categorize an unknown instance to a number of categories [72;79]. In order to use machine learning techniques, the first step requires transforming each instance into a feature representation, usually a feature vector $\boldsymbol{fv} = ((f_1, v_1), (f_2, v_2), \ldots, (f_n, v_n))$, where f_i is a feature and v_i is its corresponding value. Appropriate feature representations should capture features with high discrimination power, while the number of different features should be kept as small as possible in order to have classifiers with good generalization capabilities. The second step applies learning

algorithms to built classifiers. There are two different types of learning: supervised and unsupervised. Supervised learning refers to the process that builds classifiers by exploiting feature vectors with a known fixed number of categories, i.e., feature vectors derived from category-tagged training instances. For example, given a set of medical reports each describing a pregnancy and a birth using 200 features (e.g., patient's *weight*, *heigh*t etc), we can use supervised-learning algorithms to learn classifiers to categorize patients with high risk of emergency cesarean section. A major issue in a supervised learning task is to choose a supervised learning algorithm. In unsupervised learning, also called clustering, instances with similar feature vectors are grouped together. The similarity among feature vectors can be defined differently. For example, given a group of people, the similarity can be based on gender, or age, etc. A major issue in unsupervised learning is how to measure the similarity of two feature vectors.

Supervised Learning

The information used by supervised learning methods can be classified into two broad types: statistical information from the whole collection of instances including Naïve Bayes learning [26], decision tree [91], decision list [123], transformation-based learning [9], neural network [39], support vector machine [110], inductive logic programming [81]; and similarity among individual instances including instance-based learning [5;14]. In the following, a summarization of algorithms that have been implemented is given.

Naive Bayes Learning

Naïve Bayes (NB) learning [26] is widely used in machine learning due to its efficiency and its ability to combine evidence from a large number of features. An NB classifier

chooses the category with the highest conditional probability for a given feature vector; while the computation of conditional probabilities is based on the Naïve Bayes assumption: the presence of one feature is independent of another. The training of the Naïve Bayes classifier consists of estimating the prior probabilities for different categories as well as the probabilities of each category for each feature.

Decision Tree and Variants

A decision tree is a rooted tree where each node is either a decision node with two or more successors or a leaf node with an associated category label [91]. A decision node contains a test based on feature values. If the test has k possible outcomes, the decision node will have k branches, one associated with each outcome. Given a feature vector, searching the category to which the feature vector will be assigned is achieved via a sequence of decisions along a path of decision nodes that originates at the root and ends at a leaf node. The training of a decision tree usually applies a greedy search scheme in which the best decision test for next node is chosen at each step. For each outcome of the test, a new descendant node is created. Training instances are then sorted to leaf nodes based on the outcome of the test. If all instances associated with a leaf node are from the same category, then the node is presented as a leaf node in the resulting tree; otherwise, a test will be associated with that node later. The training of a decision tree is done if each instance in the training set has been perfectly classified to a leaf node.

A decision tree can be transformed to a set of rules. A simplified version of the resulting classifier is an ordered set of rules, where at most one rule (i.e. the first applicable rule) can be used in a classification task. Such a classifier is also termed as a decision list.

Generally in a decision list classifier, all features consist of a set of tests, and tests are ordered according to an appropriate measure that is a function of co-occurrence information of features and categories.

A different classifier with a set of ordered rules is a transformation-based learning classifier[9], where each rule is applied subsequently according to the order in a classification task. The training of a transformation-based classifier is an error-driven greedy procedure where the rule that best corrects the current errors is added at each step. Transformation-based classifiers are superior than decision lists for correcting errors made by previous rules. However, transformation-based classifiers usually have a larger set of rules than decision lists. Besides that, the number of rules that are applied to classify an instance using a transformation-based classifier is the same as the number of rules in the rule set, which is not the case for a decision list classifier.

Instance-based learning

Instance-based leaning [5;14;85] has appeared in several areas with different names: exemplar-based, case-based, and memory-based, etc. It is a form of supervised learning from instances, based on keeping full memory of training instances and classifying new instances using the most similar training instances. Instance-based classifiers can be used without training if the similarity measure between two instances is local, i.e., the similarity between two instances is totally determined by their associated feature vectors. Sometimes, instance-based classifiers include a training phase, where a set of representative cases (to reduce the number of training instances presented to the classifier) and/or a similarity measure between two instances (to include distributional

information in the similarity measure)[7] are chosen. A critical part of instance-based classifiers is the similarity measure.

Unsupervised learning -- clustering

A number of different clustering algorithms have been proposed that are more or less appropriate for different data collections and interests[50;108]. There are two different kinds of clustering approaches, hierarchical and non-hierarchical. In hierarchical approaches, clusters are arranged in a clustering tree where related clusters occur in the same branch of the tree. There are two kinds of hierarchical techniques: the agglomerative and the divisive. For a collection of n instances, the agglomerative algorithms first create n clusters (nodes) where each cluster (node) contains an instance of the collection. Then in each step, the two most similar clusters (nodes) are merged into a new cluster (node) (the merging process is recorded as the edges of the tree). The algorithms stop when only one cluster (node) is left. In contrast, divisive methods start when all instances are together and in each step a cluster is partitioned, until there are n of them. In non-hierarchical approaches, clusters are flat and the relations among clusters are undetermined. Non-hierarchical clustering algorithms are also called partitioning algorithms. For a collection of n instances, a non-hierarchical algorithm groups the collection to k clusters, with the condition that each cluster contain at least one instance and each instance belong to one cluster. Here k is given by the user or selected automatically. The similarity $\mathsf{sim}(x, y)$ of two instances x and y takes on values between 0 and 1. Similarity may be the result of subjective judgments. $\mathsf{sim}(x, y)$ can also be

[7] An instance-based classifier with a similarity measure including distributional information is also a statistics-based classifier.

computed using formulas. For instance, if *x* and *y* can be presented inside of a ball with diameter 1 in a Euclidian space, $\mathrm{sim}(x, y)$ can be computed as *1- d(x,y)* where *d(x, y)* is the distance between *x* and *y*. The similarity of two clusters can be the maximum, minimum, or average similarity between instances from these two clusters. It can also be the similarity of representatives, such as centroids, of the clusters.

2.1.4.2. Feature Representation for WSD

Machine learning of WSD classifiers requires transforming each training instance into a feature representation. Different kinds of feature representations have been exploited.

Local Co-occurring Words: Co-occurring words in the context of an ambiguous word *W* in a fixed window size are critical to WSD. For example, in the sentence "A spokesman said Healthvest has paid two of the three banks it owed interest in October", words such as *paid* and *banks* tend to indicate *interest* here holds the sense a fixed charge for borrowing money other than other senses such as a sense of concern with and curiosity about someone or something or a reason for wanting something done[8].

Local Collocations: A local collocation refers to a short sequence of ordered words. It is also important for the sense determination of *W*. For example, in the sequence "in the interest of", the sense of *interest* is the a reason for wanting something done sense of interest even though words *in*, *the*, and *of* are usually included in the stop word list for word indexing of information retrieval systems.

Derived Features: Derived features are derived from surrounding words of *W* in a window of a fixed size considering the orientation and/or distance from *W*. A derived

feature may also consist of implementing further linguistic knowledge, such as part of speech (POS) tags, semantic categories (e.g., classes in Roget thesaurus) or stemming techniques, which groups inflected forms of a root to a common feature (e.g., *discharged*, *discharging*, and *discharges* are treated as the same feature *discharg*).

2.1.4.3. Supervised WSD Methods

Several supervised learning methods have been adopted to WSD: Naïve Bayes learning [11], neural network [109], decision list [123], instance-based learning [45;86], and inductive logic programming [80]. Bruce and Wiebe [11] applied the Bayesian algorithm and chose features based on their "informative" nature. They tested their work on the *interest* corpus and achieved a precision of 79%. Towell et al. [109] constructed a WSD classifier that combined the output of a neural network that learns topical context with the output of a network that learns local context to distinguish among the senses of highly ambiguous words. The accuracy of the classifier was tested on three words, the noun *line*, the verb *serve*, and the adjective *hard*; the classifier has an average precision of 87%, 90%, and 81% respectively. The WSD system of Yarowsky [123] used the decision list method on features that consisted of both POS tags and oriented distances of the surrounding words. He claimed that the system had a precision of 99% when evaluated automatically for the accent restoration task, which is a case of the WSD problem, in Spanish and French. Ng and Lee [85] described a WSD system that uses the instance-based method with multiple kinds of features. An ambiguous term in an instance was assigned to the sense of its most similar instance in the training set in the initial version; later the sense was determined by a fixed number of the most similar instances.

---

[8] We use the gloss definition of WordNet here for senses of *interest*.

2.1.4.4. Unsupervised WSD Methods

Unsupervised WSD methods refer to WSD methods using machine learning techniques without supervision including WSD methods using unsupervised machine learning techniques and WSD methods that combine machine-readable dictionaries with raw corpora.

Clustering Analysis

Schutze [99];[100] applied clustering techniques to WSD by hierarchical clustering of word senses. He used post-hoc alignment of clusters for word senses. Schutze's results indicate that for coarse binary distinctions, unsupervised techniques can achieve results approaching a precision of around 90% for most words. Pedersen and Bruce [87] compared different clustering techniques including three clustering algorithms on WSD and showed a negative impact for rare senses. Clustering analysis has been used to overcome the data sparseness problem (see Section 2.1.8.1).

Using Machine-Readable Dictionaries and Raw Corpora

The machine-readable dictionaries alone or raw corpora alone do not provide enough information for reliable disambiguation [45;48]. Many researchers have combined machine-readable dictionaries with raw corpora for WSD. Yarowsky [124] used sense definitions as one of the options for initial sense indicators for an unsupervised WSD method. Luk [70] applied sense definitions, co-occurrence information of concepts in a small corpus. All usage examples of the sense definition were used as sense indicators in the work of Karov and Edelman [49]. Similar to Karov and Edelman's work, the system

proposed by Cho and colleagues [19] learned a set of typical usages listed in the MRD for each of the senses of an ambiguous verb using verb-object co-occurrence information that was acquired from a corpus. The system achieved an overall precision of 86.3% when evaluated on a Korean corpus.

Semantic categories have also been combined with corpora. Yarowsky [121] derived classes of words by starting with words in the same category in Roget's. A set of instances with a window size of 50 for each word in the category was extracted from the Grolier's Encyclopedia. Salient words that appeared significantly more often in the context of a category together with their weights were obtained from the set. The sense assignment was then complete using Bayes' Rule[9]. The system correctly disambiguated 92% of the instances for 12 ambiguous words.

Conceptual relations are also utilized together with corpora under the following observations: terms with certain relations tend to appear in similar contexts. For example, *summer* and the springtime sense of *spring*, can appear in similar contexts, such as "Spring is my favorite season" and "Summer is my favorite season". The corresponding type of WSD methods uses unambiguous terms that have certain relations with *W*, such as hypernymy or synonymy, in a machine-readable dictionary to derive a sense-tagged corpus automatically for use with a supervised WSD classifier. The method proposed by Leacock and colleagues [64] belongs to this type. By collecting instances of unambiguous terms from WordNet [29] that are terms associated with *W* via certain relations, such as synonymy or hyponymy, a sense-tagged corpus is automatically established for the training of a WSD classifier of *W*. An example given in their paper is

[9] A sense is represented by its associated semantic category.

the ambiguous word *suit*, where one sense has an unambiguous related term, *business suit*, and the other has an unambiguous related term *legal proceeding*. By collecting instances containing *business suit* and *legal proceeding*, a sense-tagged corpus for *suit* is automatically built by substituting these two phrases with *suit*. However, the method requires the existence of unambiguous terms with certain relations and the existence of instances of those terms in a raw corpus. Restricting their method to synonyms, direct hyponyms and direct hypernyms, they found that they could derive sense-tagged instances for about 64% of the words in WordNet from the 30-million-word corpus of the San Jose Mercury News. Milhalcea and Moldovan [77] tried to overcome these requirements by using word definitions provided by glosses in addition to close-related terms, and a very large corpus consisting of texts electronically stored on the Web. The method was tested on 20 ambiguous words (7 nouns, 7 verbs, 3 adjectives, 3 adverbs), and acquired 80,741 instances. Among 1,081 instances that were among the top ranked documents with a maximum of 10 instances for each sense, 981 were correct according to human judges. However, Agirre and Martinez [2] claimed that instances acquired through the Web are nearly useless based on the disappointing result of evaluating the corpus acquired from Web using Milhalcea and Moldovan's method, where decision list learning was the machine learning algorithm and a portion of SemCor was the test set.

Other Unsupervised WSD Methods

Different translations of senses of an ambiguous word in two languages have been used in the WSD research either through an aligned bilingual corpus or two monolingual corpora with one from each language and a machine translation method. An aligned bilingual corpus consists of two corpora that contain the same text in different languages

(for example, the Canadian Hansard, the proceedings of the Canadian Parliament which is published in both French and English). After sentence alignment, these corpora have been used for WSD by considering words with senses that translate differently across languages. Gale and colleagues [37] used a bilingual French-English corpus. For an English word *W*, the sense of *W* in a specific context was determined based on the different translations in French for the different senses of *W*. For example, *pen* in English is *stylo* in French for its writing implement sense, and *tenclos* for its enclosure sense. Kikui[53] proposed an unsupervised method that uses bilingual corpora without the alignment requirement. The method combines clustering techniques of Schutze [99;100] with a machine translation WSD method [52] and achieves a precision of about 79%.

Part-of-speech tags play an important role in the disambiguation of word senses [118] for all content words. If two senses of an ambiguous word hold different POS tags, the current state of the art POS taggers can resolve the ambiguity with a high precision. For example, the senses of *duck* in the following two sentences (a) and (b) can be disambiguated by the POS tags: *duck* is a noun in sentence (a) and a verb in sentence (b).

(a). *The duck was delicious*.

(b). *Before he could duck, another stone struck him*.

Regularities of a verb with respect to the semantic class of its arguments, called selectional preferences, are considered to be important knowledge for WSD. For example, subjects of the verb *think* tend to be human; objects of the verb *drink* tend to be a fluid. The WSD work of Resnik [94] made use of the WORDNET hierarchy and selectional preferences of verbs to disambiguate nouns. His method required that

sentences in a raw corpus be parsed so that syntactic relations such as subject-verb, verb-object, head-modifier, and modifier-head can be extracted. In a given instance of an ambiguous word *W*, examining all occurrences of nouns that have the same syntactic relation as that of W in the instance, the sense of *W* is assigned as the most commonly shared sense of those nouns and *W*. An example given in their paper is the determination of the sense of *coffee* in the phrase "drink coffee". The noun *coffee* has four senses in the WORDNET: beverage, tree, seed, or color. After extracting all occurrences of "drink XXX", such as occurrences of "drink water", "drink tea", or "drink wine", etc., the most close sense among *coffee*, *water*, *tea* and *wine* is the beverage sense.

2.1.5. Methods to Reduce Manual Annotation Cost

Although supervised WSD approaches have the drawback of requiring a sense-tagged corpus, they tend to give a higher accuracy compared to unsupervised WSD approaches [45;84;86]. As argued by Ng [84], a large sense-tagged corpus is necessary for achieving broad coverage WSD systems with high precisions. Researchers have explored intelligent methods that can reduce manual annotation cost. Generally, there are three different techniques for reducing the amount of instances that need to be sense-tagged: the bootstrapping technique, the sampling technique, and the clustering technique.

The bootstrapping technique eliminates the need for a large training set by relying on a relatively small number of sense-tagged instances of each sense for each term of interest. These labeled instances are used as seeds to train an initial classifier. Then applying the initial classifier, a larger training set is extracted automatically from the remaining untagged corpus. Repeating this process results in series of classifiers with improved precision and coverage. An early example of such an approach is due to Hearst [40]. In

his CatchWord algorithm, several occurrences of a set of ambiguous nouns are manually sense-tagged. The system automatically acquires and modifies statistical information based on newly disambiguated occurrences with a high degree of certitude. An initial set of at least 10 occurrences was indicated to be necessary, with about 20 or 30 occurrences for high precision. A subsequent similar bootstrapping technique is described in Yarowsky [124]. There are a variety of options for selecting seed instances in his paper: words in dictionary definitions, a single collocation defined in the dictionary for each sense, or manually acquired collocations from a corpus. The method was evaluated on binary sense disambiguation for 12 words and achieved a precision of about 95%.

There are several sampling techniques proposed in the literature. Observing that if an instance's classification is uncertain given current annotated instances then the instance is likely to contain unknown information which is useful for classifying similar instances in the future, Lewis and Gale [67] proposed a method called sequential sampling, and Engelson and Dagan [27] proposed a method called committee-based sampling. Both sampling techniques are used for statistics-based WSD classifiers. Fujii et al. [35] proposed a selective sampling method for instance-based classifiers. The method selectively samples a smaller-size effective subset from a given example set for use in WSD.

Schutze's unsupervised WSD methods using clustering analysis can also be regarded as methods to derive sense-tagged corpora when adding a post-hoc alignment phase which assigns senses to clusters [87;99;100].

2.1.6. Implementation of WSD Systems

As mention by Ide and Veronis [45], methods for WSD have evolved largely independently of particular applications in the recent past. Obviously, WSD research should be of benefit to machine translation, information retrieval and information extraction. For example, the correct translation of *pen* to *stylo* or *tenclos* in French depends on the correct sense disambiguation of *pen* in the context, i.e., its writing implement sense or its enclosure sense. However, machine translation systems have not incorporated recent WSD methods except the system of Kikui [53] (See Section 2.2.4.3.3). WSD research also has an impact on information retrieval. For example, it is desirable to eliminate documents containing the word *aids* that are associated with "hearing aids" when searching for the disease "AIDS". However, the majority of attempts to improve information retrieval using WSD were unsuccessful. Krovetz & Croft [61] and Sanderson [98] determined that a WSD system can improve information retrieval only if: queries are short, there are no rare senses, there is a highly accurate WSD system, sense distinctions are coarse-grained, and sense definitions cross grammatical boundaries. There is not much research that combines WSD with information extraction systems since most information extraction systems are applied on very specific domains, where domain specific words or terms are usually not ambiguous, and general English words do not play an important role in the systems.

2.1.7. Evaluation and Performance

The evaluation of WSD methods is impeded by the lack of large-scale gold standard sets [56] and systematic evaluation methods. There are currently two different evaluation methods [86]: attempting to apply WSD to all the content words of texts [78;118], or

restricting the evaluation to a small number of selected words [11;65;83]. SemCor is usually used for the former, while the DSO corpus, the *interest* corpus, and the *line* corpus, etc., are often used for the latter.

Two basic measures used in evaluating WSD classifiers are precision, the ratio of the number of instances that are tagged correctly to the number of instances that have been tagged, and recall, the ratio of the number of instances that are tagged correctly to the total number of instances.

It is generally believed that the best dictionary-based WSD performance can be achieved by mixing all kinds of knowledge from MRDs [85;104] as illustrated in the first Senseval competition [57], which was a competition of WSD systems. A problem with such hybrid systems is that they are difficult to implement.

Co-occurring words and collocations are used in almost all machine learning WSD methods. There is no agreement on the preference of window sizes, i.e., the number of neighboring words that should be included as sources for deriving features. It is also generally believed that nouns require a larger window than verbs [45]. Obviously, large values of window sizes capture dependencies at longer range but also dilute the effect of the words closer to the term. Leacock et al. [65] used a window size of 50, while Yarowsky [123] argued that a small window size of 3 or 4 had better performance. A small window size has an advantage of requiring less system space and running time.

There is no agreement on the performance of supervised learning methods for WSD. Leacock and colleagues [65] showed that various supervised learning algorithms tended to perform roughly the same when given the same evidence. Mooney [80] reported that

Naïve Bayes learning gave the best performance on disambiguating the *line* corpus among seven learning algorithms tested. Ng [83] reported that performance of instance-based classifiers were comparable to Naïve Bayes classifiers on the DSO corpus. Yarowsky [123] stated that decision list classifiers had at least as good performance as Naïve Bayes classifiers with the same evidence and also had the advantage of easy interpretation, easy modification and easy implementation.

2.1.8. Other Issues

2.1.8.1. Data Sparseness

Unlike other machine learning tasks that have a limited number of features, machine learning methods on free text need to handle a very large number of features, along with the zero-frequency of co-occurrences of features [119]. Frequency-based and information-retrieval-based methods have been applied to select features that best discriminate one category from others [87;100]. The zero-frequency problem can be solved using smoothing techniques, class-based methods or similarity-based methods. Smoothing techniques [21] reevaluate co-occurrence statistics by assigning "zero probability" to some non-zero values [85;123]. Class-based methods [10;88;93] cluster words into classes of similar words, so that one can estimate words' co-occurrences from the average co-occurrences of the classes to which these words belong. Some authors [10;88;100] derived classes from the distributional properties of the corpus itself, while some others [121] used semantic categories from machine-readable dictionaries to define classes. Similarity-based techniques [19;24;25;49] exploit the clustering idea but without grouping words to fixed classes. Each word is modeled by its own set of similar words derived from statistical data extracted from corpora rather than fixed classes.

2.1.8.2. Sense Definition and Sense Granularity

Almost all the WSD work assumes a set of predetermined senses for an ambiguous word. However, it is a nontrivial task to determine a set of senses for a word because sense is an abstract concept frequently based on subjective and subtle distinctions in topic, dialect, collocation, etc [75]. Various approaches to derive a set of predetermined senses have been used in the WSD work, including i) senses defined in every-day dictionaries [23;66], ii) automatic or handcrafted clusters of dictionary senses [11;70], iii) thesaurus categories [121], iv) translations in another language [37], v) automatically induced clusters [99;100], and vi) handcrafted lexicons [75].

Using senses defined in machine-readable dictionaries has the advantage of an automatic derivation of a set of senses. However, the sense division in an MRD is listed along grammatical lines and frequently too fine-grained for the purpose of WSD when related to NLP applications. For example, Sanderson [98] and Krovetz [59] studied the impact of WSD for information retrieval systems independently and both found that coarse-grained sense distinctions that cross grammatical boundaries may be appropriate for information retrieval system. Using thesaurus categories such as those listed in Roget's Thesaurus is also problematic. Yarowsky [121] reported that 3 out of 12 nouns have uses not listed in Roget's Thesaurus, while some uses that a native speaker might consider holding a single sense are often encoded in several Roget's categories. Using translations in another language suffers the incompleteness problem, i.e., many ambiguities are preserved in the target language (e.g. French translation of *interest*). A WSD system built using automatically induced clusters as senses is useless for NLP applications without a post-hoc sense alignment process. The automatic clustering of word senses that cross

grammatical boundary seems to be a reasonable choice in handling word senses for NLP applications [59;59;98;98].

2.1.8.3. One sense per collocation and One sense per discourse

As a consequence of "Similar context implies similar senses", two hypotheses have been studied in the 1990s: one sense per discourse [36] and one sense per collocation [122]. In their experiments with WSD, Gale, Church and Yarwosky [36] observed a strong relationship between discourse and meaning. They proposed a hypothesis: one sense per discourse -- when a word occurs more than once in a discourse, all occurrences of that word will share the same meaning. They conducted an experiment using 9 ambiguous words and a total of 82 pairs of concordance lines for those words, and showed that 94% occurrences of ambiguous words from the same discourse have the same meaning. One sense per collocation was observed and quantified [122] with 97% correct for adjacent content words. The measures were reported on coarse-grained distinction of senses (i.e., distinguish *bank* as a bank of a river or as a financial bank). Both hypotheses are weaker for fine-grained distinction of senses (i.e., a financial bank sense of *bank* will split to several senses such as a depository financial institution, savings bank, or the funds held by a gambling house). Krovetz [60] reported 67% when using fine-grained distinction of senses on two manually tagged corpora: Semcor [29] and DSO [83], where WordNet was used as the sense inventory. Martinez and Agirre [73] reported 70% for one sense per collocation using the same corpora and sense inventory as Krovetz. Both hypotheses have been used for WSD in the general English domain [124].

2.2. Comparison of Our Work with Previous Work

The most obvious difference between our work and previous work is that we define our WSD work in the biomedical domain instead of in the general English domain. The resulting WSD system is intended to be implemented within real-world NLP systems in the biomedical domain.

Besides this difference, the proposed WSD method differs from traditional WSD work that is based on machine-readable dictionaries. First, it applies a biomedical machine-readable knowledge base instead of machine-readable dictionaries in the general English domain. Unlike the work of Agirre and Rigau [4] who used conceptual relations in machine-readable dictionaries alone and considered the disambiguation of all content words, our method combines machine-readable dictionaries with machine learning techniques, and considers only ambiguous biomedical terms. Our method is not designed for all content words since most content words in the biomedical domain are not ambiguous or not important for the corresponding NLP applications. Unlike the work of Mihalcea and Moldvan [76] that used statistics from the Web that may contain rich genres of contexts, our method extracts instances from free-text databases on which NLP applications are employed. Unlike previous work that uses automatic derivation of sense-tagged corpora [64;77], our method applies clustering analysis to examine the quality of the derived corpora.

Unlike previous comparison studies [28;80], we compared the performance of WSD classifiers using different combinations of feature representations, machine learning algorithms, window sizes, different sets of ambiguous words, with or without the existence of noise in the training set.

Our implementation of decision list learning uses two sets of tests that distinguish features that occur with only one sense from other features while traditional implementations of decision list learning algorithm do not distinguish them and apply smoothing techniques to avoid zero-frequency of co-occurrence of features with senses. Our implementation has a better performance than traditional implementations when there is no noise in the training set. Traditional implementations of Naïve Bayes learning do not distinguish rare senses from majority senses in the training set. We split these two and propose a hybrid supervised learning algorithm that combines a Naïve Bayes classifier with an instance-based classifier.

We do not use sampling or bootstrapping techniques to reduce the amount of sense-tagged instances needed for WSD, but apply clustering analysis. Bootstrapping needs to decide which kind of supervised learning methods to use beforehand, and the resulting classifier is difficult to implement. Selective sampling favors instance-based classifiers; sequential sampling and committee-based sampling favor statistical classifiers. Those sampling techniques require human experts be interactive with the processes many times. Additionally, the sense-tagged corpus may not be suitable for deriving classifiers using other learning methods. It is not known beforehand how many instances are required to be annotated using bootstrapping techniques and sampling techniques; and also it is not easy to estimate the human effort. The clustering method proposed in this dissertation is designed to build WSD classifiers for real-world applications. The human effort (how many instances can be affordably annotated for each ambiguous term) can be estimated in advance, and the clustering analysis will then determine the final clustering based on the estimate.

The clustering algorithm presented in this dissertation is different from traditional clustering algorithms. Researchers mostly concentrate on developing clustering algorithms that can obtain optimal clusters but disregard the requirement of the speed and space of the algorithms. In addition, existing clustering algorithms are not designed to handle mixed instances, i.e., some instances are tagged and some instances are un-tagged.

Chapter 3. Resources and NLP Systems in the Biomedical Domain

3.1. Machine Readable Knowledge Base: the UMLS

The goal of the UMLS is to overcome retrieval problems caused by differences in terminologies and the scattering of relevant information across many databases by integrating different electronic biomedical terminologies into one concept-oriented knowledge base. It contains three knowledge sources: the Metathesaurus (META), the Specialist Lexicon, and the Semantic Network.

The META provides a uniform, integrated distribution format for over 60 biomedical vocabularies and classifications, and links many different names for the same concepts. Each distinct concept has been assigned a unique concept identifier (CUI). Concept names corresponding to the same concept are assigned the same CUI. For instance, *abdominal neoplasm* and *tumor of abdomen* are two different concept names with the same CUI *C0000735*.

The Specialist Lexicon contains syntactic information for many terms, component words, and English words, including verbs, which do not appear in the META. The Specialist Lexicon abbreviation list contains 10,410 unique (AW, FF) pairs in the 2000 version of the UMLS, where AW is an abbreviation and FF is the corresponding full form. Some of them are general English abbreviations, for instance, *adm* for *admission*.

The Semantic Network contains information about the types or categories (e.g., “Disease or Syndrome”, “Virus”) to which all META concepts have been assigned and the

permissible relationships among these types (e.g., "Virus" causes "Disease or Syndrome").

3.2. Free-text Databases: MEDLINE and Clinical Data Repository

MEDLINE is the premiere bibliographic database of the National Library of Medicine (NLM) which contains 11 million references to journal articles in life sciences with a concentration on biomedicine. Each entry contains the citation information to the corresponding journal article, including authors, title, sources, often an abstract, and the index information that facilitates the MEDLINE search.

The New York Presbyterian Hospital (NYPH) Clinical Data Repository (CDR)[42] is a collection of electronic medical records. It provides a location for the storage and retrieval of data placed by health care professionals or computer applications. The repository contains narrative data as well as coded data. The narrative data contains reports from the domains of discharge summary, radiology, neurophysiology, pathology, GI endoscopy, Ob/Gyn, cardiology, surgery, and so forth.

3.3. NLP Systems

We developed our WSD system with the goal of integrating it with real-world NLP systems in the biomedical domain. Two systems were involved in this research: MedLEE[33] and MetaMap[7]. In the following, we give an overview for each.

3.3.1. MedLEE

MedLEE was designed as a general information extraction and encoding language processing system within the clinical domain. It was initially developed for chest

radiographs, and has since been expanded to the domains of mammography, radiology reports, pathology reports, echocardiography, electrocardiography and discharge summaries. A number of evaluations of the system were performed within the domains of chest radiography, mammography and discharge summary reports [32;41;58] which demonstrated that MedLEE was effective in identifying specific clinical conditions, and MedLEE was effective for a clinical application that resulted in improving the quality of patient care.

The current version of MedLEE [31] has five functional components together with several corresponding knowledge components. Figure 4 shows the different components. The oval components are knowledge bases; the other components are the programming engines. The preprocessor uses a lexicon (Lex), a list of abbreviations (Ab), a list of section names (Sec), and disambiguation contextual rules (Cru) for lexical lookups. A brief summary of each functional component is presented below.

The preprocessor performs lexical lookups in order to recognize and categorize words and phrases using a lexicon and a list of local contextual disambiguation rules. The preprocessor also identifies sentences and abbreviations. For instance, the output of the preprocessor for "spleen was enlarged" is the following structure, [*(spleen, bodyloc, spleen), (was, vbe, be), (enlarged, cfinding, enlarged)*], where *bodyloc*, *vbe* and *cfinding* are semantic categories and *spleen*, *be* and *enlarged* are target forms in the lexicon. The parser uses a grammar to identify the structure of the sentence and to generate an intermediate structure based on grammar specifications. The grammar is a list of rules based on semantic and syntactic co-occurrence patterns. The output for "spleen was

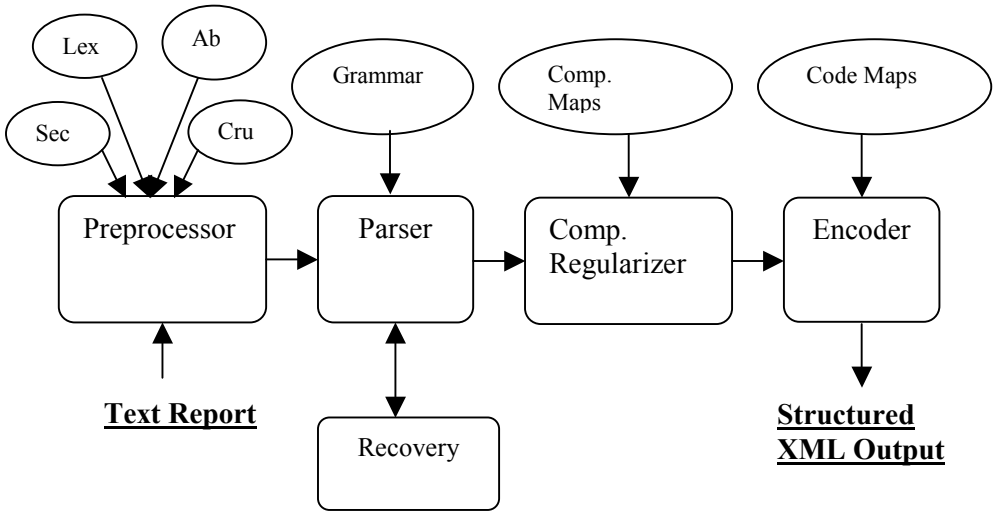


Figure 4. An overview of components in the MedLEE

enlarged" in this stage is the following, [*problem, enlarged, [bodyloc, spleen], [certainty, be]]*. The compositional regularizer uses a table of structural mappings to compose individual words into phrases. After composing, the output of the parser for "spleen was enlarged" is [*problem, enlarged spleen, [certainty, be]]*. The encoder maps words and phrases into controlled vocabulary terms if applicable. The intermediate output is then mapped into XML. The final XML output for spleen was enlarged is *<problem v= "splenomegaly">< certainty v="high certainty"/></problem>* where the controlled term for *enlarged spleen* is *splenomegaly* and the controlled term for *be* is *high certainty*. The recovery component increases sensitivity by using alternative strategies to structure the text if the initial parsing effort fails.

3.3.2. MetaMap

MetaMap is a highly configurable program that maps biomedical text to concepts in the META. The program was initially developed to improve retrieval of bibliographic material such as MEDLINE citations. It has been applied to several applications including concept indexing, terminology discovery, genomic information, knowledge discovery, and the NLM Indexing Initiative project [7;43;96;102;113].

The performance of MetaMap was tested by Henny and Klein on 100 randomly chosen MEDLINE abstracts, with a total of 13,426 words. MetaMap produced over 7,000 concepts with an average precision (the ratio of the number of correct concepts to total concepts) of 94.35%, and an average coverage (the ratio of correctly mapped words to total words) of 63.55%.

MetaMap consists of five functional components as well as several knowledge components. Figure 5 shows these components: the oval components are knowledge bases; the other components are programming engines. Both the syntactic parser and variant generator use the Specialist Lexicon. In addition the variant generator uses a synonym set (Syn), and a list of abbreviations (Ab). A brief summary of each component is presented below.

The syntactic parser parses arbitrary text into (mainly) simple noun phrases (i.e., noun phrases without preposition attachment) using the SPECIALIST minimal commitment parser. For example, the phrase *inventory of interpersonal problem* is determined to have two noun phrases: *inventory* and *of interpersonal problem*, where words with part of speech tags such as prepositions, conjunctions and determiners are normally ignored in

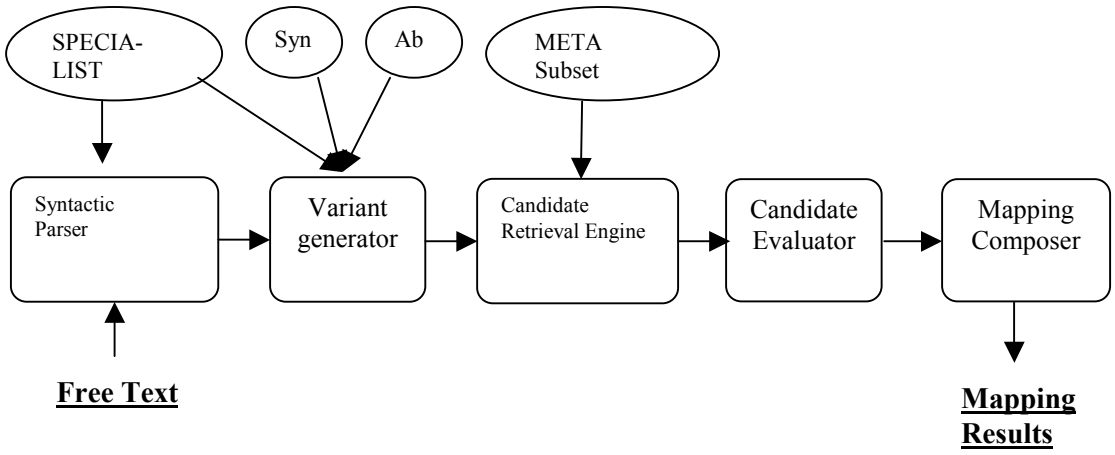


Figure 5. An overview of components in the MetaMap.

the processing. For example, *of* in *inventory of interpersonal problem* will be ignored. The variant generator generates variants for each phrase word using knowledge in the Specialist Lexicon and a synonym set. Variants of a phrase word include itself together with all of its abbreviations, synonyms, derivational variants, meaningful combinations of these, and finally inflectional and spelling variants. For example, variants of *inventory* include *inventory*, *invent*, *inventories*, and *invents*. The candidate retrieval engine retrieves all META strings containing at least one of the variants. Each META string is evaluated by the next component against the input text by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping using several metrics. The candidates are then ordered according to their mapping strength. The final component is a mapping composer where complete mappings are constructed by combining candidates involved in disjoint parts of the phrase, and the strength of the complete mappings is computed. The highest scoring complete mappings represent MetaMap's best interpretation of the original phrase. For

example, the best mapping for *interpersonal problem* is C0814588, where the preferred term is an exact match.

Three subsets of META have been created for different levels of requirement for the accuracy of the program:

- Strict: is appropriate for semantic processing where the highest level of accuracy is needed, and consists of about half of the English META strings.
- Moderate: is appropriate for term processing where input text should not be divided into simple phrases but considered as a whole.
- Relaxed: is appropriate for browsing.

3.4. Ambiguity in the Biomedical Domain

3.4.1. Ambiguity in the MedLEE System

The lexicon of MedLEE categorizes medically relevant words and phrases, and specifies their target forms. A lexical entry (s, t) for a term in the semantic lexicon consists of the semantic category (s), and the target form (t). For example, the term *abdominal* belongs to the body location (bodyloc) category and the target form is *abdomen*. Some terms may be associated with multiple (s, t) pairs, i.e., they are ambiguous. For example, the term *head* has two (s, t) pairs: *head* is associated with (bodyloc, head) in the phrase "head scan" and (region, head region) in the phrase "the femoral head". There are 269 words (or phrases) associated with multiple (s, t) pairs in the current MedLEE lexicon (the March 2002 version). MedLEE uses a set of handcrafted rules that are based on contextual information to achieve disambiguation for some ambiguous words. For example, the

following rule is needed to disambiguate *hr*: if it is preceded by one, per or 1, *hr* stands for hour; otherwise, *hr* stands for heart rate.

3.4.2. Ambiguity in the UMLS

There are two different kinds of ambiguity presented in the UMLS: conceptual and semantic. Conceptual ambiguity refers to the ambiguity caused by terms denoting multiple concepts such as the term *discharge*, while semantic ambiguity refers to the ambiguity caused by terms having multiple semantic categories such as concepts belong to the semantic category "Organic Chemicals" most likely also belong to the category "Pharmacologic Substance". Some terms are semantically ambiguous but not conceptually ambiguous. For example, most concept names that are "Organic Chemicals" are not conceptually ambiguous. Most, but not all conceptually ambiguous terms are also semantically ambiguous, i.e., they name concepts from different semantic categories. There are some terms that are conceptually ambiguous, but not semantically ambiguous. For example, two concepts denoted by the term *acetate*, C0000975 and C0000979, are organic chemicals. With the widespread use of abbreviations, the number of terms that are conceptually ambiguous but not semantically ambiguous increases. For example, the term *IBD* in MEDLINE abstracts denotes at least three different diseases: inflammatory bowel disease, infectious bursal disease, and ischemic brain disease.

Aronson considered some conceptual ambiguities as unnecessary, such as ambiguity caused by terminology sources that have contextual strings. For example, *regular* which is listed as an ambiguous string in the AMBIG.LUI with two senses: C0369532 (i.e., *regular insulin*) and C0205272 (i.e., qualitative modifier *regular*). However, the sense C0369532 of *regular* is due to the context string *insulin* in the LOINC terminology.

There are 187,943 (out of 797,359) concepts that possess multiple semantic categories in the UMLS[10]. There were 4,547 conceptually ambiguous terms that represented 11,178 concepts in the UMLS ambiguous term table AMBIG.SUI, with an average ambiguity of 2.46[11]. Johnson [46] investigated the semantic ambiguity of a semantic lexicon that was based on the UMLS and discharge summaries, and proposed a set of preference rules to reduce the semantic ambiguity. For example, in the discharge summary domain, chemical concepts occur only under the semantic category *chemicals viewed functionally* instead of under *chemical viewed structurally*. After applying his preference rules to the derived semantic lexicon, occurrences of entries with multiple semantic types were reduced from 9.41 to 1.46 per cent in discharge summaries. Rindflesch and Aronson [95] considered the conceptual ambiguity of the UMLS and proposed to use neighboring concepts' semantic categories to resolve the ambiguity. They conducted a preliminary study, and found that a manually crafted set of rules based on the semantic categories of neighboring concepts successfully resolved conceptual ambiguity around 80% of the time. Aronson and colleagues [6] proposed that machine learning techniques could be used to derive rules instead of the manually crafting process. However, there is no published study of this approach according to our knowledge.

3.4.3. Types of Ambiguity

The ambiguity of biomedical terms as partially stated by Roth and Hole [97] can be classified into four different types:

[10] In the dissertation, the UMLS is the 2001 version of the UMLS if there is no specification.
[11] We use the AMBIG.SUI table that lists ambiguous concept names using string identifier. A different table AMBIG.LUI that lists ambiguous concept names using term identifiers is not used in the dissertation

1. General biomedical ambiguous terms – terms that are identical but with different biomedical meanings, for instance, the term *discharge* can mean either the discharge procedure as in "prior to discharge" or the discharge substance as in "bloody vaginal discharge". The senses of general biomedical ambiguous terms are usually a subset of the senses found in general language (e.g. *discharge*). However, this is not always true (e.g., the word *girdle* does not contain the body region sense of *girdle,* as in "pelvic limb girdle", in the online general English lexicon WordNet[29]).

2. Identical biomedical abbreviations – abbreviations that have multiple full forms, for instance, *APC* abbreviates *activated protein c, adenomatosis polyposis coli, adenomatous polyposis coli, antigen presenting cell, aerobic plate count, advanced pancreatic cancer, age period cohort, alfalfa protein concentrated, allophycocyanin, anaphase promoting complex, anoxic preconditioning, anterior piriform cortex, antibody producing cells,* and *atrial premature complex,* etc., in MEDLINE abstracts. Abbreviations contribute a large portion of ambiguous biomedical terms. Many clinical findings, diseases and procedures have been abbreviated[8;18] because brevity is favored in the biomedical domain writing.

3. Identical biomedical and general world terms-- terms that have senses from both the general world domain and the biomedical domain. For instance, the term *lead* can mean a chemical substance as in "lead shield overlies the pelvis" or an electronic lead as in "a single lead pacemaker" or the verb lead as in "these lead us to the right colic vein", and the word *add* can be an abbreviation for *attention deficit disorder*.

4. Contextually ambiguous terms - terms that have different semantic interpretations depending on different contexts. For example, chemical terms in the context of laboratory tests (e.g. *iron* in *triple sugar iron test*) denote test items, whereas in the context of medication items (e.g. *iron* in *iron drops with fluoride*), they represent medication drugs.

3.4.4. A WSD Test Collection

Realizing the need for a WSD evaluation set in the biomedical domain, Weeber and colleague manually built a WSD test collection that consists of 50 highly frequent ambiguous UMLS concept names from the 1998 version of MEDLINE. Each of the 50 ambiguous terms has 100 ambiguous instances (i.e., senetences) randomly selected from MEDLINE. For a total of 5,000 instances, 11 subjects attended the tagging process, of which 8 completed 100% of the 5,000 instances, 1 completed 56%, 1 completed 44%, and the remaining one completed 12% of the instances. The assessment of the evaluation results found 12 of the 50 terms were problematic: subjects disagreed with each often. For example, there were four senses for word *adjustment* as listed in the following.

M1 - Adjustment <1> (Individual Adjustment) [inbe, Individual Behavior]

M2 - Adjustment <3> (Adjustment Action) [ftcn, Functional Concept]

M3 - adjustment <5> (Psychological adjustment) [menp, Mental Process]

None - None of the Above

Among 10 subjects that sense-tagged the highlighted *adjustment* in the following sentence, 4 chose M1, 3 selected M2, 2 picked M3, and 1 subject tagged as an undefined sense, None.

"These variables accounted for 62% of the variance (58% adjusted) in **adjustment** when adjustment at diagnosis was controlled".

3.5. Summary

All previous WSD work in the biomedical domain was done manually. As we saw in Section 3.4., a large-scale WSD system is required for NLP applications. However, manual generation of WSD rules is very time-consuming and costly. In addition, maintenance of rule sets becomes increasingly difficult over time. Moreover, hand-coded rules are often incomplete and perform poorly when extended to a broader domain. In this dissertation, we propose a two-phase WSD method for NLP applications in the biomedical domain based on information gathered from a machine-readable knowledge base, the UMLS, and two large-scale free-text databases, MEDLINE and CDR. The method applies NLP techniques, supervised machine learning analysis, and clustering analysis, as well as expert knowledge when needed.

Chapter 4. Methods

Our WSD method contains two phases for the construction of a WSD classifier for an ambiguous term *W* in the biomedical domain. The first phase derives a sense-tagged corpus for *W*, STC(*W*), from a collection of free-text documents in the biomedical domain. In the first phase, a preliminary sense-tagged corpus for *W* is automatically derived from a free-text collection using the UMLS. Clustering analysis is then applied optionally on STC(*W*) to check the quality of the automatic derived corpus after transforming each instance in STC(*W*) to a feature representation. If the corpus has good quality, it becomes STC(*W*). Otherwise, expert annotation is required for clusters with a relatively large size but containing no sense-tagged instances, and STC(*W*) consists of instances in expert-annotated clusters besides instances in the automatic derived corpus. In the second phase, each instance in STC(*W*) is transformed into a feature representation and then a supervised learning algorithm is implemented to derive a WSD classifier. Note that features for clustering analysis and features for supervised learning may be different. In the following, we first describe the methodology to automatically derive sense-tagged instances using the UMLS; feature representations of the context are presented next; the clustering algorithm is then discussed; and the automatic construction of WSD classifiers is described last.

4.1. Automatic Derivation of Sense-Tagged Corpora

In this section, we discuss the automatic derivation of a sense-tagged corpus for an ambiguous term *W* using conceptual relations. In the following, the definition of conceptual relatives is given in Section 4.1.1. The relations in the UMLS are presented in

Section 4.1.2, the derivation of the representative set for a UMLS concept is described in Section 4.1.3. We then describe the automatic derivation of conceptual relatives for a given term in Section 4.1.4; and Sections 4.1.5 and 4.1.6 describe the automatic derivation of sense-tagged corpora using the UMLS.

4.1.1. Definition of Conceptual Relatives

Many machine-readable dictionaries (MRD) contain a rich set of relations that link senses. For example, all nouns in WordNet [29], which is a handcrafted MRD, are organized into one conceptual network through the hypernymy relation. For a term *W*, we define a term that has a relation R with a sense *S* of *W* in a conceptual network as a **conceptual relative** of *W* via the associated sense *S* and the associated relation R. For example, the word *summer* is a conceptual relative of the word *spring* via the sense *the springtime sense of spring,* and the sibling relation (since *summer* and *spring* share a common parent: *the season of the year)*.

4.1.2. Relationships in the UMLS [12]

Since the META is organized by concept, terms with the same concept identifiers are synonyms. For instance, *congestive heart failure* is a synonym of *biventricular heart failure* since they have the same concept identifier (C0018802). Relations other than synonymy relations are listed in the MRREL table. There are 9,524,132 entries in MRREL. Among them, 9,518,798 were derived directly from the source vocabularies. The remaining 5,334 entries are relationships between different sources that were created during META construction. There are 9 different relationship types:

- Broader (RB): a broader relationship, e.g., C0007222 (i.e. *cardiovascular disease*) has an RB relation with C0018802 (i.e., *congestive heart failure*) in WHO97 and MTH.

- Narrower (RN): a narrower relationship, a reverse relationship of RN (i.e., Broader relationship).

- Other related (RO): a relationship other than synonymous, narrower, or broader, e.g. C0018800 (i.e., *Cardiomegaly*) has an RO relation with C0018802 (i.e., *congestive heart failure*) in MTH.

- Like (RL): the two concepts are similar or "alike", e.g. C0000755 (i.e., *abnormal hard tissue formation in pulp*) and C0011434 (*secondary dentin*) have an RL relation in SNOMED.

- Parent (PAR): a parent relationship in a META source vocabulary, e.g., C0018802 (i.e., *congestive heart failure*) is a parent of C0007193 (i.e., *congestive cardiomyopathy*) in MeSH.

- Child (CHD): a child relationship in a META source vocabulary, a reverse relationship of the parent relationship.

- Sibling (SIB): a sibling relationship in a META source vocabulary, e.g., C0013274 (i.e. *patent ductus arteriosus*) is a sibling of C0018802 (i.e., *congestive heart failure*) in the source vocabulary COSTAR.

---

[12] Refer to http://www.nlm.nih.gov/research/umls for descriptions of Sources such as WHO97, MTH, AOD99, MTH2001 etc. The relationships discussed here refer to the UMLS 2001 version; the relationships in the UMLS 2002 version are slightly different from the description here.

- AQ : an allowed qualifier for a concept in a META source vocabulary, e.g., C0005768 (i.e., *in blood*) is a qualifier for C0018802.

- QB: can-be-qualified by a concept in a source vocabulary of the META, a reverse relationship of the qualifier relationship.

However, since relations in the UMLS were mostly derived from different source vocabularies, the definition of relationship types may not be consistent among different source vocabularies. For example, two concepts may have multiple relationship types defined in the MRREL table. For example, the concepts C0004015 (i.e. *aspartic acid)* and C0085845 (i.e., *aspartate)* have a parent relation and a broader relation from source vocabulary AOD99; they have a narrower relation from source vocabulary MSH2001; while in source vocabulary LNC10o, they have an RO relation. A concept may have a relation with itself. For example, the concept C0022709 *angiotensin converting enzym*e has an RO relation with itself in source vocabularies CSP2000 and LNC10o.

4.1.3. The Representative Set of a UMLS Concept

For each UMLS concept C, we gather all unambiguous English concept names of C. Because concept names with a term status "suppressed" are incomplete[13], we exclude them; in addition, because abbreviations are highly ambiguous, we exclude those identified as abbreviations by our UMLS abbreviation extraction program, which will be described in Section 5.1. The concept names are normalized by changing to lower-case, removing symbols such as *NOS* in *Cerebrospinal fluid, NOS*, removing some patterns such as parenthetical expressions *(CK)* in *Creatine kinase (CK)*, or *CK* – in *CK - Creatine*

[13] The term status "suppressed" indicates that a string is less useful and problematic.

kinase, and substituting some punctuation marks with blanks. The resulting strings with length greater than 4 form the representative set of C (note that strings with length less than or equal to 4 appear frequently as abbreviations in documents even though they may not be ambiguous by themselves). For example, the representative set of C0009392 is {*cell growth inducer myeloid, colony stimulating factor, colony stimulating factors, inducer myeloid cell growth, mgi 1, mgi 1 protein, myeloid cell growth inducer*}.

4.1.4. Derivation Methods

Let *W* be an ambiguous term and let the set *SEN*={*S1*, *S2*, …, *Sn*} be its *n* senses. Let CUI_{Si} be the concept identifier that represents the sense *Si*. We denote the set {CUI_{S1}, CUI_{S2}, …, CUI_{Sn}} as SCUI(*W).* For example, the SCUI(*W*) for the following four senses of the abbreviation *CSF* is {C0007806, C0009392, C0072454, C0893357}:

- CSF_1: *Cerebrospinal fluid* (C0007806),
- CSF_2: *Colony stimulating factor* (C0009392),
- CSF_3: *Cytostats factor* (C0072454),
- CSF_4: *Competence and sporulation factor* (C0893357).

4.1.4.1. Establishing Conceptual Relative Sets

For each sense *Si* of *W*, concepts that have direct relations (i.e., concepts with CUIs that co-occur with CUI_{Si} in the MRREL table) with *Si* consist of conceptual relatives. We consider concepts with relationship types “RB”,”RN”,”RO”, “RL”, “PAR”, “CHD”, “SIB”, and exclude concepts with relationship types “QB” or “AQ” since they are qualifier relationship types, have high frequency, and provide little sense disambiguation

information. The concept identifier of each concept is put in the relative concept identifier set of *W* (RCUI(*W*)) with its associated sense *Si* and the relations. We consider that each concept has a synonymy relation with itself, and put each CUI_{Si} of *W* in RCUI(*W*) with its associated sense *Si* and a relationship type synonymy but disregard relations among different senses of *W* in the MRREL table. For example, C0020255 (i.e. *hydrocephalus*) and C0007806 (i.e. CUI_{S1} of *CSF*) have a RO relation in MRREL; therefore C0020255 is put in RCUI(*CSF*) with its associated sense CSF_1 and RO relation. All strings in representative sets of concepts in RCUI(*W*) with the associated senses and relations consist of conceptual relatives for *W*.

4.1.4.2. Automatic Generation of a Sense-tagged Corpus Using Synonyms

For each sense *Si*, all instances containing strings from the representative set of CUI_{Si} are extracted from free-text databases. For each instance, the string from the representative set is replaced by *W*; and the sense of *W* in the instance is annotated as *Si* and is put into $STC_1(W)$. For example, the sense-tagged instance for *CSF* generated from **Instance 1** will be **Instance 1'.**

Instance 1. *The Optic Neuritis Treatment Trial continues to generate information and controversy on the visual and neurologic outcome and treatment of optic neuritis. At the same time, other researchers explored cerebrospinal fluid parameters in multiple sclerosis, treatment of experimental optic neuritis, corticosteroid treatment of multiple sclerosis, and variations and mimickers of optic neuritis.*

Instance 1'. CSF_1 *| The Optic Neuritis Treatment Trial continues to generate information and controversy on the visual and neurologic outcome and treatment of optic neuritis. At the same time, other researchers explored CSF parameters in multiple sclerosis,*

treatment of experimental optic neuritis, corticosteroid treatment of multiple sclerosis, and variations and mimickers of optic neuritis.

4.1.4.3. Automatic Generation of a Sense-tagged Corpus Using Conceptual Relatives in the Context

We assume that multiple occurrences of *W* hold the same sense in MEDLINE abstracts, i.e., one sense per abstract. The context for acquiring disambiguation knowledge is the whole abstract.

There are several steps to generate a sense-tagged corpus for *W* using conceptual relatives. First, all abstracts that contain *W* are extracted from MEDLINE. The second step is to identify conceptual relatives in each abstract. The third step is to assign senses to abstracts with identified conceptual relatives based on certain criteria.

Identification of Conceptual Relatives in an Instance

A program, CRMap, is used to identify conceptual relatives. CRMap consists of the following phases: preprocessing, exact-string matching, UMLS-Specialist normalization matching, and stem-normalization matching.

In the preprocessing phase, we remove parenthetical expressions that contain a capitalized term with fewer than six characters. This is based on the observation that parenthetical expressions containing a short capitalized term inside are usually abbreviation type parenthetical expressions. The punctuations are replaced by blanks. This phase also changes the text to lower case. As an example, the text "The influence of prednisone on S-angiotensin-converting enzyme (S-ACE) activity was examined…" is

changed to “the influence of prednisone on s angiotensin converting enzyme activity was examined …”.

The three matching phases are processed subsequently. All three matching phases match conceptual relatives of the longest possible length. The matching phases differ in whether they require normalization or not, and if so, the normalization method used. In the exact-matching phase, conceptual relatives are used without normalization, while in the UMLS-Specialist normalization matching phase, CRMap normalizes each word in the conceptual relative set and abstracts and maps it to its base-form in accordance with the Specialist Lexicon LRAGR table if applicable. In the stem-normalization matching phase, CRMap use Porter-stemmer [89] to normalize each word to its stem.

For example, in the following abstract **Instance 2** that contains the abbreviation *CSF*, we have three conceptual relatives identified where each is associated with the sense CSF_1 (i.e., *cerebrospinal fluid*): *hydrocephalus*, *spinal cord*, and *brain.*

Instance 2. *The **brain** (CSF_1_SIB) from an infant with a cystic occipital mass present at birth is examined in serial section. The occipital mass proved to be a rhombic roof ventriculocele. Within the posterior fossa, it was bound to an occipital lobe encephalocele which issued as a diverticulum of the left lateral ventricle through a microgyric cortical defect in the territory of the left posterior cerebral artery. The posterior medial aspects of both cerebral hemispheres were herniated downward into the widened tentorial gap. Craniolacunae were prominent on the inner aspect of the skull. The aqueduct and central canal of the **spinal cord** (CSF_1_SIB) were widely dilated, although the lateral ventricles were collapsed. It is suggested that **hydrocephalus** (CSF_1_RO) secondary to obstruction to flow of CSF through the rhombic roof entrained a sequence of events giving rise to the rhombic roof ventriculocele and causing occlusion*

of the posterior cerebral artery and subsequent diverticulation of the lateral ventricle through an infarcted region of the posterior-medial hemisphere.

Besides CRMap, the MetaMap program can also be used to identify concepts that have relations with senses of *W*.

Criteria for Sense Assignment

For each abstract, since all occurrences of *W* in the abstract hold the same sense *S* based on the one sense per abstract assumption, we call *S* the sense of *W* in that abstract. For an abstract that has conceptual relatives identified by CRMap, the sense of *W* in the abstract is the majority vote of associated senses of the identified conceptual relatives; if there is a tie, we randomly choose one of the tied senses. For example, the sense of *CSF* in **Instance 2** is CSF_1 (i.e. *cerebrospinal fluid*).

4.2. Feature Representation

As we stated before, appropriate feature representations should capture features with high discrimination power, while the number of different features should be kept as small as possible. The main body of WSD research about feature representation has been pursued in the general English domain and it is agreed that co-occurring words and local collocations are appropriate features. Also it is observed that topical nouns require a larger window size than other words. We use a feature vector $\boldsymbol{fv} = ((f_1, v_1), (f_2, v_2), \ldots, (f_n, v_n))$ to represent an instance, where f_i is a feature and v_i is the number of occurrences of f_i in the instance**.** In the experiment chapter, we will describe our experimental comparisons

of different combinations of features and window sizes[14] for supervised learning algorithms and our choice of features for clustering analysis.

4.3. Clustering Analysis

The purpose of clustering analysis is two-fold: first, it is used to automatically check the comprehensiveness of the automatic derived sense-tagged corpora derived using the UMLS; and secondly, it reduces the human annotation cost by grouping similar contexts together when expert supervision is required. The input to the clustering analysis contains sense-tagged instances as well as instances that contain *W* but cannot be sense-tagged using relations in the context. Instances are grouped together based on certain similarity measure. If a corpus is determined to be comprehensive by some subjective criteria, then expert annotation is not required. Otherwise, expert annotation is performed. A large size corpus can be derived when assigning each un-tagged instance the majority sense in that cluster.

Our design of clustering analysis takes account of the number of instances of each term that can be affordably manually annotated by an expert as well as sense-tags of sense-tagged instances.

We use the following steps to define our clustering task:

Feature selection and similarity: based on one sense per collocation observation, instances that share the same collocations are similar to each other and should be grouped into one cluster. For example, an instance of *white* containing "…in the white matter and…" and an instance of white containing "…of the white matter can…" have the same

[14] We used Hashtable to represent feature vector in our experiments.

collocation ("the white matter"). They share the same sense of white. Instances with similar neighboring words are also considered similar to each other. For example, an instance of *white* containing "…cortical gray and *white* matter…" and an instance of white containing "…cortical *white* and gray matter …" have the same neighboring words ("cortical", "and", "gray", "matter"). They share the same sense of *white*. However, stop features (i.e. features appear frequently in the text disregarding senses and terms such as "the" and "in", etc.) may contaminate the similarity among instances and should be excluded from the feature representations. For example, an instance containing "…and *white* matter in…" and an instance containing "…and *white* men in" may consider similar if we include "and" and "in" as features. During our implementation, we removed stop features. Detailed feature representations and similarity measures will be discussed in the experiment chapter.

Clustering criteria: since there is noise in the input (i.e., the sense assignment of some instances may be incorrect when using related terms in the context for the assignment), we allow sense-tagged instances in a cluster to have different senses but the majority of them must have the same sense as a percentage over a certain threshold (which is a function of the noise in STC_2) among sense-tagged instances in that cluster. The number of final clusters should also be less than a different threshold (which is a function of the affordable human supervision cost).

Clustering algorithm: our algorithm is composed of several sequential clustering iterations where the order of clusters presented to each iteration is randomized. A set of

Initialization : $C = \{c_1, c_2, c_3, ..., c_N\}$, where $c_i = \{x_i\}$

$t = t_1$

Interation :

While ($|C| \geq n_1$ and $t \geq t_M$)

{

Randomize {1,2,..., $|C|$} to {r1, r2, .., r$|C|$}

$C' = \{c_{r1}\}$

For i = 2 to $|C|$

{

Find a cluster c' in C' that is most similar to c_{ri}, where c' satisfies the condition

that the percentage of the majority sense when merging c' and c_{r1} is over n_2

If the similarity of c' and c_{ri} is larger than t

{

Replace c' in C' by $c_{r1} \cup$ c'

}

Else

{

C' = C'$\cup\{c_{ri}\}$

}

}

If ($|C| - |C'| < 5$)

{

Assign t to the next similarity threshold in T

}

C = C'

}

Return C

Figure 6. The clustering algorithm

decreasing similarity threshold values is bought into compliance with clustering iterations.

Assume there are N instances in STC(W), let them be $x_1, x_2, x_3, ..., x_N$. Let n_1 be the number of clusters that are affordable to be manually annotated. Let n_2 be the percentage threshold of the majority sense in a cluster. Let $T = \{t_1, t_2, t_3, ..., t_M\}$ be a decreasing similarity threshold vector.

The algorithm is illustrated in Figure 6. We first initialize a clustering C that has N clusters, where each cluster contains one instance in STC(W). For each similarity threshold value t in T, there are several iterations. In each iteration, we first randomize the order of clusters in C (i.e., {1, 2, …, |C|}) to a new order {r1, r2, …, r|C|}, and initialize a clustering C' with the first cluster (i.e., c_{r1}) in C. For each cluster c_{ri} in C, we find a cluster c' which is most similar to c_{ri} and satisfies the majority sense criteria (i.e., the percentage of the majority sense is over n_2) when merging c' to c_{ri}. If the similarity between c' and c_{ri} is larger than the similarity threshold value t, we update c' in C' to the merged cluster of c' and c_{ri}. Otherwise, we add the cluster c_{ri} to C' (i.e., the number of clusters in C' is incremented by 1). After each iteration, if the difference of the number of clusters in C' and C is less than 5, we assign the similarity threshold variable the next value in T if applicable. We then assign C the new clustering C' and begin a new iteration. The iteration stops when the number of clusters in C is less than n_1 (i.e., the number of clusters that are affordable to be manually annotated) or there is no more similarity threshold value available in T.

Let C_1 be the clusters that have at least one instance from the derived sense-tagged corpus. If the number of instances in C_1 is over 90% of the total number of instances and there are no clusters with a relatively large number of instances such as more than 5 but

having no sense-tagged instances, then the derived sense-tagged corpus can be considered comprehensive. A large size sense-tagged corpus of *W* can be derived when assigning each un-tagged instance the majority sense of the corresponding cluster if applicable.

The input to the clustering algorithm may contain only un-tagged instances. In this case, the purpose of clustering analysis is to reduce human annotation cost since only one or two instances from each cluster are presented to human experts for sense-tagging. Additionally, the iteration may stop when the number of clusters has not dropped under n_1 but there is no more similarity threshold value available in *T*.

4.4. Automatic Construction of WSD Classifiers

For an ambiguous word W, once we have a sense-tagged corpus, any robust supervised machine learning algorithm can be used to derive a WSD classifier for W. Robust here means that the algorithm can tolerate noise and rare senses. Implementations of several supervised learning algorithms have been developed using PERL language and will discussed in detail in the experiment chapter.

Chapter 5. Automatic Derivation of Gold Standard Sets and Summary of Evaluation Sets

In this chapter, we first present a method that automatically extracts an abbreviation knowledge base from the UMLS. We then demonstrate how to derive the gold standard sense of an abbreviation *AW* for an instance that defines *AW* using a parenthetical expression. Information about each evaluation set is also provided.

5.1. UAExtractor: A Method to Extract an Abbreviation Knowledge Base from the UMLS

5.1.1. Background

In the META, the names that contain abbreviations are treated as synonyms of the names that contain their full forms, and therefore they are assigned the same concept identifier. For instance, *ERV* and its full form *expiratory reserve volume* are both listed as one of the names of the same concept (i.e., C0015326). Some concept names actually include the abbreviation together with the full form, e.g. *expiratory reserve volume (ERV)* and *ERV - expiratory reserve volume*. *ERV* by itself is also listed as an abbreviation in the Specialist abbreviation list. However, not all abbreviations in the META have a corresponding entry in the Specialist abbreviation list and vise versa. For instance, *TTTS,* which stands for *twin to twin transfusing syndrome* in the META has no entry in the Specialist abbreviation list while *APT*, which stands for *aminopropylisothiuronium*, is in the Specialist Lexicon abbreviation list but not in the META.

5.1.2. Abbreviation Extraction Method

The extraction program is based on manual observation of a training set. The training set contained 36,899 concept names, which were concept names in English whose concept

identifiers contained the prefix C000[15]. The output generated by the extraction program is a list of (AW, FF) pairs, where AW is an abbreviation and FF is the corresponding full form. The program handles the following three cases.

Case 1: An abbreviation and the phrase containing its full form are connected by a dash.

In this case, the abbreviation appears on the left side of the dash and the phrase on the right side. The full form can be the whole phrase, e.g. *AV* - *aortic valve* or a sub-string of the phrase, e.g. *AV – arteriovenous fistula* or *AV - abnormal atrioventricular connection* (the full form of an abbreviation is underlined).

Case 2: An abbreviation and its full form are included in a parenthetical expression.

In this case, the abbreviation appears inside the parentheses or immediately to the right. In the former case, the full form is a rightmost sub-string of the phrase to the right of the parentheses, e.g. *insertion of intrauterine device (IUD)*. In the latter case, the full form is a whole phrase included inside the parentheses, e.g. *CAD (coronary artery disease)*.

Case 3: An abbreviation and its full form occur in different concept names associated with the same concept identifier.

There are two types of abbreviations defined in this case. The primary type occurs when the abbreviation and its full form occur as two different concept names associated with the same concept, e.g. *ADP* and *adenosine diphosphate*. The derived type is derived from the primary type. For instance, we derive two abbreviation pairs *(abd, abdominal)* and *(cav, cavity)* from a primary type abbreviation pair (*approach through abd cav, approach through abdominal cavity).*

[15] The 2000 version of the UMLS was used.

An abbreviation knowledge base containing pairs of (*AW*, *FF*) is constructed using the following several steps:

1. Extracting a list of (*AW*, *FF*) pairs from the META using the extraction program;
2. Merging results with the SPECIALIST abbreviation list;
3. Removing subsumed pairs: a pair (*AW*, *FF2*) is a subsumed pair of (*AW*, *FF1*) if each word in *FF2* can be matched to an equivalent portion (either an equivalent word or a full form of that word) in *FF1*.

Two words are considered to be equivalent if they are the same or have the same base form in the Specialist Lexicon. For instance, in the following, (b) and (c) are two subsumed pairs of (a): *ischaem* is an abbreviation of *ischaemia*, and *ischemia* and *ischaemia* have the same base form in the Specialist Lexicon.

(a). (*AMI, acute mesenteric ischaemia*)

(b). (*AMI, acute mesenteric ischemia*)

(c). (*AMI, acute mesenteric ischaem*)

The extraction program was evaluated to have an accuracy of 97.5% and a recall of 96% using the 2000 version of the UMLS[68].

5.2. Automatic Derivation of the Gold Standard Sense for an Abbreviation

Utilizing the fact that authors usually define abbreviations when they are first introduced in documents, the gold standard sense of *AW* for an instance that contain the parenthetical

pattern "*FF* (*AW*)" can be determined automatically, where *AW* is an abbreviation, and FF is the associated full form. Two different methods can be used to derive the sense of *AW* in an instance containing "*FF* (*AW*)*"*. One method actively matches *AW* with *FF* using a pattern-matching method, and then associates *FF* with its corresponding sense using a concept mapping method. For example, a pattern-matching method, based on regularities that authors form abbreviations, can detect that *cerebrospinal fluid* is a full form of *CSF* in Instance 3 (since letters *C* and *F* are initials and *S* appears in the middle of the first phrase word). The sense of *CSF* in Instance 3 can be known if a concept mapping method, for example, MetaMap, detects that *cerebrospinal fluid* is a concept name of the UMLS concept C0007806. The other method applies an abbreviation knowledge base to check the existence of a known full form of *AW*, *FF* (or its variants), at the left side of the pattern "(*AW*)", as described in the following paragraph. Since a WSD task usually begins with a set of predefined senses, the novelty introduced by the first method makes the task not well defined, i.e., the sense of *FF* that is found by the matching method may not have a correspondence in a given set of senses of *AW*. We use the second method to derive the gold standard set while leaving the first method to several studies presented in Chapter 7.

Instance 3. *After a brief summary of current views on the origin of cerebrospinal fluid (CSF) and the processes underlying its elaboration, the author discusses studies of isolated chorid plexus in extracorporeal perfusion….*

If an instance contains a pattern "*FF* (*AW*)*",* where *FF* is associated with a definition of *AW* in a knowledge base, then we consider the gold standard sense of *AW* in the instance to be the associated sense of *FF.* After determining the correct sense of *AW*, the instance is automatically modified by replacing the pattern *FF (AW)* with *AW*, and then put in the

gold standard set of *AW*. For instance, **Instance 3** of CSF is modified to **Instance 3'** with the gold standard sense attached at the beginning (separated using the sign "|"), where CSF_1 is the sense identifier for cerebrospinal fluid. Only modified abstracts are used for further processing. The sense of the beginning of each abstract is used for evaluation purposes to determine correctness, but not used by the disambiguation method itself.

Instance 3'. *CSF_1 |After a brief summary of current views on the origin of CSF and the processes underlying its elaboration, the author discusses studies of isolated chorid plexus in extracorporeal perfusion….*

5.3. Evaluation Sets

We used three sets of ambiguous terms, A, B and C for the experiments. Set A contains 35 frequently occurring ambiguous abbreviations in the medical reports where the gold standard set for each abbreviation was derived automatically from MEDLINE. Set B contains 38 general ambiguous terms used in the WSD project of National Library of Medicine (NLM), where the gold standard set was determined manually by Weeber and his colleagues [114]. Set C contains 4 ambiguous terms, i.e., *cold*, *discharge*, *lead*, and *dressing*, in the clinical domain, where the gold standard set has been derived manually using human experts.

5.3.1. Set A

Set A contains frequently appearing ambiguous three-letter abbreviations in medical reports, where frequency information is from a collection of medical reports and sense definitions of those abbreviations are from an abbreviation knowledge base extracted from the UMLS 2001 version.

We used a collection of medical reports to generate frequency information. The collection consisted of reports of patients admitted during 1998 at NYPH in the following domains: discharge summary, radiology, neurophysiology, pathology, GI endoscopy, Ob/Gyn, cardiology, and surgery. The number of occurrences of each three-letter capitalized string was derived from the collection. The abbreviation knowledge base that contains pairs of (*AW*, *FF*) was derived using the 2001 version of the UMLS, where *AW* is an abbreviation and *FF* is the corresponding full form. We kept only those pairs where the *AW* i) was listed as an ambiguous term in the UMLS ambiguous terms table, ii) had multiple full forms, iii) appeared more than 100 times in the collection of medical reports, and iv) *FF* was a UMLS concept name.

For each abbreviation *AW* in Set A, we collected all MEDLINE abstracts that contained *AW* inside a parenthetical expression. The gold standard set of *AW* was derived using the method described above. We also derived a set of gold standard instances for *AW* from the collection of medical reports using the same method.

There were 35 abbreviations that met the criteria for the study. The average ambiguity for the set, i.e., the average number of senses, was 3.8. The ambiguity here refers to the ambiguity captured by the UMLS. Table 2 shows the detailed information for a few representative abbreviations (*AW*), where SID is the assigned sense identifier and CUI is the UMLS concept identifier of the corresponding full form (see Appendix A for detailed definitions for each abbreviation). For example, the two full forms of *BSA* are *body surface area* and *bovine surface area*, which have been assigned sense identifiers BSA_1 with the associated CUI (i.e., C0005902) and BSA_2 with the associated CUI (i.e. C0036774).

We extracted 80,681 abstracts from MEDLINE that had an occurrence of *AW* that was inside a parenthetical pattern "(*AW*)", where *AW* is an abbreviation from Set A. Among them, 70,764 abstracts had gold standard senses identified for the corresponding abbreviations using our method and consisted of the gold standard sets. The average ambiguity for abbreviations in Set A was 3 in the gold standard sets. There were 460 instances with gold standard senses determined from the collection of medical reports. Table 3 shows the information about each abbreviation: the number of senses defined in the UMLS (DS), the number of senses with instances in the gold standard set (ES), the number of instances in the gold standard set (GSS), the majority sense (MJS), the number of instances in the majority sense (MGSS) and it percentage, and the number of instances extracted from the collection of medical reports (CMR).

5.3.2. Set B

Set B contained 38 ambiguous terms that were considered to be non-problematic in the study of Weeber et al [114] (refer to Section 3.4.4). We downloaded the WSD test collection from the Web[16]. Instances in the test collection were sentences. We transformed sense definitions in the collection using the 2001 version of the UMLS since Weeber et al. used the 2000 version of the UMLS for the sense definitions in their study (see Appendix B for the detailed sense definitions). In addition, an occurrence of an infected variant of an ambiguous word was considered to be an ambiguous occurrence in the test collection of Set B. For example, an occurrence of *extracts,* such as in the sentence "Extracts were analyzed with 1H and 31P NMR spectroscopy and metabolite peaks were quantified using an external standard", was considered to be an occurrence of

[16] See http://skr.nlm.nih.gov

extraction. In our study, only occurrences of an exact ambiguous term were considered as occurrences of that term and were included in the gold standard set. Table 4 summarizes the statistics of the 38 non-problematic terms using senses present in the 2001 version of the UMLS: NS is the number of senses presented in the test collection, NGSS is the number of instances with senses in the UMLS, and NONE is the number of instances with the sense "None" (i.e., there are no correspondent UMLS concepts for senses of those instances).

5.3.3. SET C

The four ambiguous words in Set C were *cold*, *lead*, *discharge*, and *dressing*[43;97;101]. Table 5 shows the detailed information about each word. The senses listed in the table were defined by two subjects (X and Y, both hold an MD degree) who referred to the MedLEE semantic lexicon, the META, and instances from the corpora. The gold standard set for each term consists of 50 instances that were manually tagged by the same subjects using the following method. First, subject X tagged 50 instances for each of the two words c*old* and *lead*; and subject Y tagged 50 instances for each of the other two words. Then subject Y checked the result that was tagged by subject X; and subject X checked the result that was tagged by subject Y. Because of the coarse granularity of the sense definitions and instances from medical domain, two subjects agreed with each other totally.

AW	SID	CUI	Full Form
ACE	ACE_1	C0001044	acetylcholinesterase
	ACE_2	C0022709	angiotensin converting enzyme
	ACE_3	C0050385	doxorubicin cyclophosphamide
	ACE_4	C0108844	doxorubicin cyclophosphamide etoposide
	ACE_5	C0286421	amsacrine cytarabine etoposide
	ACE_6	C0304721	adrenocortical extract
	ACE_7	C0473028	antegrade colonic enema
APC	APC_1	C0003315	antigen-presenting cells
	APC_2	C0032580	adenomatous polyposis coli
	APC_3	C0033036	atrial premature complexes
	APC_4	C0085171	aphidicholin
	APC_5	C0809732	activated protein c
ASP	ASP_1	C0038013	ankylosing spondylitis
	ASP_2	C0003431	antisocial personality
	ASP_3	C0003993	asparaginase
	ASP_4	C0004015	aspartic acid
	ASP_5	C0052546	aspartylglycine
	ASP_6	C0085845	aspartate
BSA	BSA_1	C0005902	body surface area
	BSA_2	C0036774	bovine serum albumin
CSF	CSF_1	C0007806	cerebrospinal fluid
	CSF_2	C0009392	colony stimulating factors
	CSF_3	C0072454	cytostatic factor
	CSF_4	C0893357	competence and sporulation factor
EMG	EMG_1	C0004903	exomphalos macroglossia gigantism
	EMG_2	C0013839	electromyography
	EMG_3	C0180677	electromyographs
	EMG_4	C0393125	electromyogram
IBD	IBD_1	C0021390	inflammatory bowel diseases
	IBD_2	C0022104	irritable bowel syndrome
MAS	MAS_1	C0016065	mccune albright syndrome
	MAS_2	C0025048	meconium aspiration syndrome
	MAS_3	C0451273	macandrew alcoholism scale
PVC	PVC_1	C0032624	polymer vinyl chloride
	PVC_2	C0151636	premature premature complex
	PVC_3	C0280556	cisplatin cyclophosphamide etoposide
RSV	RSV_1	C0035236	respiratory syncytial virus
	RSV_2	C0086943	rous sarcoma virus
VCR	VCR_1	C0042679	vincristine
	VCR_2	C0182936	videocassette recorder
	VCR_3	C0526312	vanadyl ribonucleoside complex

Table 2. The detailed information for a few abbreviations, where AW is an abbreviation, SID is the sense identifier, and CUI is the UMLS concept identifier.

AW	DS	ES	GSS	MJS	MGSS	(%)	CMR
ACE	7	6	5,856	ACE_2	5,820	99.4	-
ANA	3	2	896	ANA_2	843	94.1	3
APC	5	5	2,310	APC_1	1,356	58.7	392
ASP	6	5	141	ASP_6	60	42.6	-
BPD	4	3	906	BPD_2	465	51.3	-
BSA	2	2	3,162	BSA_2	2,808	88.8	-
CAD	5	3	3,325	CAD_1	3,294	99.1	-
CAT	5	3	36	CAT_1	34	94.4	-
CML	2	2	3,350	CML_1	3,178	94.9	1
CMV	4	4	4,944	CMV_1	4,887	98.8	8
CPI	3	3	72	CPI_2	59	81.9	-
CSF	4	3	10,771	CSF_1	9,962	92.5	6
CVA	2	1	226	CVA_1	226	100.0	-
CVP	3	2	587	CVP_3	581	99.0	-
DIP	3	2	112	DIP_3	81	72.3	4
DOB	3	2	2	DOB_1	1	50.0	-
DVT	2	2	1,598	DVT_1	1,584	99.1	2
EMG	4	3	3,770	EMG_3	2,036	54.0	-
FDP	5	4	431	FDP_3	382	88.6	-
HSV	2	2	3,479	HSV_1	3,398	97.7	2
IBD	2	1	1,149	IBD_1	1,149	100.0	1
LAM	5	4	183	LAM_1	103	56.3	-
LDH	2	2	3,390	LDH_1	3,389	100.0	-
MAC	9	6	862	MAC_3	535	62.1	1
MAS	3	2	112	MAS_2	81	72.3	-
MCP	6	5	461	MCP_5	185	40.1	1
PCA	9	6	1,553	PCA_3	507	32.6	7
PCP	5	4	2,225	PCP_4	1,071	48.1	-
PEG	2	2	70	PEG_1	52	74.3	1
PSA	3	3	3,227	PSA_2	3,215	99.6	27
PVC	3	2	571	PVC_1	473	82.8	-
RSV	2	2	1,954	RSV_1	1,335	68.3	-
SLE	2	2	6,772	SLE_2	4,887	97.7	3
TPN	2	2	1,623	TPN_2	1,621	99.9	-
VCR	3	2	638	VCR_1	634	99.4	-
TOTAL	132	104	70,764	NA	60,292	85.2	459

Table 3. Statistical information for Set A. DS is the number of defined senses in the UMLS, ES (GSS) are the number of senses (instances) presented in the gold standard set, MJS is the majority sense, MGSS is the number of instances holding the majority sense, CMR is the number of instances extracted from the collection of medical reports.

WORD	DS	ES	GSS	MJS	MGSS	NONE
ASSOCIATION	3	1	98	None	98	98
COLD	6	5	98	$COLD_1$	86	5
CULTURE	3	2	54	$CULTURE_2$	49	0
DEGREE	3	3	65	$DEGREE_1$	58	5
DEPRESSION	3	2	90	$DEPRESSION_1$	75	15
DISCHARGE	3	3	75	$DISCHARGE_2$	52	23
ENERGY	3	2	96	$ENERGY_2$	95	0
EXTRACTION	3	3	39	$EXTRACTION_1$	29	6
FAT	3	3	95	FAT_2	68	26
FIT	3	2	60	None	55	55
FLUID	3	1	81	$FLUID_1$	81	0
FREQUENCY	3	2	100	$FREQUENCY_1$	94	6
GANGLION	3	2	58	$GANGLION_2$	51	0
GLUCOSE	3	2	100	$GLUCOSE_1$	91	0
GROWTH	3	2	100	$GROWTH_2$	63	0
IMPLANTATION	3	3	100	$IMPLANTATION_2$	81	2
INHIBITION	3	3	100	$INHIBITION_2$	98	1
JAPANESE	3	3	100	$JAPANESE_2$	73	21
LEAD	3	3	99	None	71	71
MAN	4	4	100	MAN_1	58	8
MOLE	4	3	9	$MOLE_1$	6	2
NUTRITION	4	4	53	$NUTRITION_1$	26	0
PATHOLOGY	3	3	95	$PATHOLOGY_2$	80	1
PRESSURE	4	2	100	$PRESSURE_1$	96	4
REDUCTION	3	3	88	None	77	77
REPAIR	3	3	98	$REPAIR_1$	51	31
RESISTANCE	3	2	100	None	97	97
SCALE	4	2	98	$SCALE_2$	65	33
SECRETION	3	2	88	$SECRETION_2$	87	0
SEX	3	2	99	SEX_2	84	0
SINGLE	3	2	100	$SINGLE_2$	99	0
STRAINS	3	3	100	$STRAINS_2$	92	7
SURGERY	3	2	99	$SURGERY_2$	97	0
TRANSIENT	3	2	97	$TRANSIENT_1$	97	0
TRANSPORT	3	3	90	$TRANSPORT_1$	89	1
ULTRASOUND	3	2	100	$ULTRASOUND_1$	84	0
WEIGHT	3	3	51	None	29	4
WHITE	3	3	100	$WHITE_2$	49	10
TOTAL	122	97	3273	NA	2731	609

Table 4. Statistical information for Set B. NONE is the number of instances holding the sense “None”. Refer to Table 3 for notations.

WORD	Sense ID	Definition CUI		Examples	GSS
Cold	1	C0009264	Low temperature	The patient feels better with cold water to drink.	12
	2	C0009443	Common cold, disease	He stated he has noted a cold with positive sputum and a runny nose for about two to three months.	9
	3	C0010412	Cold therapy, application	A cold cup biopsy	13
	4	N/A	Not active	A cold nodule	8
	5	C0234192	Feeling cold, cold sensation	At these times her finger will become pale and feel cold for as long as one hour or two.	8
	6	C0719425+N/A	A brand name, or a name of something	Dr. Cold	0
Discharge	1	C0030685	The administrative process	Discharge options including the plan to have the patient move with his sister were reviewed.	40
	2	C0563526	The electrical conduction	A single sharply contoured discharge was seen in the right posterior quadrant.	0
	3	C0012621 C0600083	A substance that is emitted or released	There was no purulent discharge.	10
Dressing	1	C0518459 C0152053	The process of putting on clothes	He is independent with dressing.	11
	2	C0278286 C0013119	The clean or sterile coverings	The wound was covered with a sterile dressing.	39
Lead	1	C0181586	Electrical conductors used in obtaining electrocardiographs or in pacemaker functions.	A single lead left-sided pacemaker is seen in good position.	40
	2	C0023175 C0373667	Metal Pb	No intra-nuclear inclusions typical of lead toxicity are identified.	0
	3	N/A	(Verb) direct, guide, tend	These findings lead to mild segmental spinal stenosis.	10
	4	N/A	The top position, the principal	Lead medical transcriber	0

Table 5. The detail information for Set C.

Chapter 6. Experiments

The experiments were designed to evaluate our two-phase method using three sets of ambiguous terms presented in the previous chapter.

We first present several studies we conducted for the second phase of our method, i.e., the automatic construction of WSD classifiers using supervised learning techniques. As we know, there is no agreement on the preferred feature representation, the suitable window used to extract features, and the best supervised learning algorithm for supervised WSD. Given an ambiguous term *W* and a sense-tagged corpus for *W*, how to construct a supervised WSD classifier for *W*? In Section 6.1, we describe our comparison study of supervised WSD classifiers using Sets A and B. Since there is noise in the derived sense-tagged corpus, we compared the noise tolerance of different supervised learning algorithms, which is presented in Section 6.2.

We then assess several hypotheses we proposed for our method to answer the following questions:

What kinds of terms can use our method to automatically derive WSD classifiers with a reliable precision (i.e., without two optional components (clustering analysis and expert annotation) in the first phase of our method)? And what kinds of terms require expert annotation?

Hypothesis 1. Our method can be used to automatically derive WSD classifiers for abbreviations in MEDLINE with a set of known full forms.

Hypothesis 2. WSD classifiers for abbreviations, which are trained on sense-tagged instances derived from MEDLINE, can also be used to disambiguate instances in the clinical domain.

Hypothesis 3. Our automatic extraction of sense-tagged instances can also be applied to derive sense-tagged instances for a majority of ambiguous UMLS biomedical terms.

Hypothesis 4. The derived WSD classifiers achieve a high precision for ambiguous UMLS biomedical terms without closely related senses provided there are enough instances.

Hypothesis 5. Clustering analysis can reduce human annotation cost dramatically.

Hypotheses 1 and 2 were proposed for abbreviations, and the proofs are presented in Section 6.3. Hypotheses 3, 4, and 5 were proposed for ambiguous general biomedical terms, and the proofs are given in Section 6.4.

6.1. Comparison Study of Supervised WSD Classifiers

There are several comparison studies in the literature about supervised WSD classifiers as described in Section 2.1.7. However, most of them had only one variable regarding to either supervised learning algorithm or feature representation. We describe a comparison study of WSD classifiers with four variables including type of ambiguous terms, feature representation, supervised learning algorithm, and window size.

6.1.1. Background about the Evaluation of Supervised Classifiers

The estimation of the performance of a supervised classifier presupposes that one has decided upon a gold standard set to which the performance measure computing will be

applied. There are several methods that provide unbiased estimation of the performance. The most popular one is the holdout method that splits the gold standard set into a training set and a test set. The classifier is built using the training set and tested using the independent test set. However, in real-world cases, the number of instances in the gold standard set is limited. There is a tradeoff between the number of instances in the training set and the number of instances in the test set. Using most data for training may yield a good classifier but the performance estimation is not persuasive; and using a small amount of data for training may yield a poor classifier. An alternative method is the leave-k-out cross-validation method that is free of the dilemma associated with the holdout method. The method repeatedly assesses the performance of classifiers trained on N-k instances and tested on the remaining k instances, where N is the total number of instances in the gold standard set and k is a positive integer. The most popular choice of k are 1 (leave-one-out) and N/10 (ten-fold).

6.1.2. Feature Selections

Six different feature representations were studied for a given window size n, which will be referred as “a”, “b”, “c”, “d”, “e”, and “f” respectively. Each word was normalized using the Specialist Lexicon, and all numbers were unified to the string XXX. Four feature representations (i.e., “a”, “b”, “c”, and “e”) depended on window sizes; while feature representations “d” and “f” were not functions of window sizes. Let “…w_{Ln}…$w_{L2}w_{L1}$W$w_{R1}w_{R2}$…w_{Rn}…” be the context of consecutive words around the term *W* to be disambiguated. Features refer to this context as follows.

- Representation "a" contains all words with their corresponding oriented distances within the window, i.e., Ln/w_{Ln},…, $L2/w_{L2}$, $L1/w_{L1}$, $R1/w_{R1}$, $R2/w_{R2,}$… , and Rn/w_{Rn}, where L is for left, R is for right, and the number is for the distance.

- Representation "b" contains all words with their corresponding orientations within the window, i.e., L/w_{Ln},...,L/w_{L2}, L/w_{L1}, R/w_{R1},$R/w_{R2,}$..., and R/w_{Rn}.

- Representation "c" contains all words within the window, i.e., w_{Ln},...,w_{L2}, w_{L1}, w_{R1}, $w_{R2,}$..., and w_{Rn}.

- Representation "d" contains all words with their corresponding orientation within a window of size 2 and three nearest two-word collocations, i.e., L/w_{L2}, L/w_{L1}, R/w_{R1}, R/w_{R2}, $L2L1/w_{L2}\_w_{L1}$, $L1R1/w_{L1}\_w_{R1}$, and $R1R2/w_{R1}\_w_{R2}$.

- Representation "e" combines features in representations "c" and "d".

- Representation "f" combines features in representation "d" and all words in the context except *W*.

For example, the above representations for CSF in **Instance 4** with a window size 3 are shown in Table 6.

Instance 4. *At the same time, other researchers explored CSF parameters in multiple sclerosis, treatment of experimental optic neuritis, corticosteroid treatment of multiple sclerosis, and variations and mimickers of optic neuritis.*

FP	Features	Example (window size = 3)
a	Words with oriented distance within the window	*L3/other, L2/researcher, L1/explore, R1/parameter, R2/in, R3/multiple*
b	Words with orientation within the window	*L/other, L/researcher, L/explore, R/parameter, R/in, R/multiple*
c	Words within the window	*other, researcher, explore, parameter, in, multiple*
d	Three collocations, oriented words within a window size 2	*L/researcher, L2L1/researcher_explore, L/explore, L1R1/explore_parameter, R/parameter, R1R2/parameter_in, R/in*
e	Features in c and d	*L/researcher, L2L1/researcher_explore, L/explore, L1R1/explore_parameter, R/parameter, R1R2/parameter_in, R/in, other, researcher, explore, parameter, in, multiple*
f	Features in d and all other words	*L/researcher, L2L1/researcher_explore, L/explore, L1R1/explore_parameter, R/parameter, R1R2/parameter_in, R/in, at, the, …, and, variation, mimickers*

Table 6. Six options of feature representation, where FP stands for feature representation.

We did not use part of speech information as features since POS taggers in the general English domain have been shown inappropriate for documents in the biomedical domain and there were no POS taggers trained specifically in the biomedical domain [13]. We did not use semantic categories as features because there are no broad-coverage semantic lexicons available in the biomedical domain. We investigated using the META as a lexicon but found that the semantic classification for many terms were problematic[34].

6.1.3. Supervised Learning Algorithms

Five different supervised learning algorithms were implemented including Naïve Bayes learning, traditional decision list learning, instance-based learning, our implementation of decision list learning, and our mixed supervised learning. The first three algorithms have been introduced in Section 2.1, and our implementations are presented as following:

Naïve Bayes Learning: we used the Witten-Bell discounting technique[119] to avoid the zero probability in the algorithm. Witten-Bell discounting is based on a simple intuition about zero-frequency events: the probability of seeing a zero-frequency feature is estimated by the probability of seeing a feature for the first time. Let N be the occurrences of all features in the training set, T be the number of different features appearing in the training set, and Z be the number of different features that have zero-frequency in the universe. The frequency of unseen features is $\frac{T}{Z} \times \frac{N}{(N+T)}$. However, Z is not known in the WSD problem. We used $\frac{T}{100 \times (N+T)}$ as the frequency of unseen features by assuming $Z = 100 \times N$.

Traditional Decision List Learning: We used the algorithm that was implemented by Yarowsky [123]. Each individual feature consists of a test. All tests are ordered according to their log-likelihood ratios: $\log(\frac{Occu(s,f)}{Occu(f) - Occu(s,f)})$, where s is the majority sense that co-occurs with f, Occu(f) is the number of occurrences of f, Occu(s,f) is the number of occurrences of f appearing in instances of W that are associated with the sense s. The default test returns the majority sense[17]. For features (f) that co-occur with only one sense, a smoothing factor 0.1 is added to the total occurrences of f.

Instance-based Learning: Our implementation of instance-based learning is a k-nearest-neighbor algorithm, where the weighted majority sense of the k nearest neighbors and the weight of a neighbor is the rank of its similarity among similarities of k nearest

[17] However, since our algorithm will be used in a circumstance that the majority sense in the training set may not be the majority sense in the universe, the default test returns a sense randomly choosing from majority senses in Section 6.2.

neighbors. For simplicity, we used the normalized inner-product of two feature vectors as the similarity measure and chose k as 3.

In **our implementation of decision list learning**, features that co-occur with only one sense are separated from others. Two sets of tests are derived during the learning. The first set consists of features that co-occur with only one sense and are ordered according to the following formula: $\log(\frac{Occu(f)}{Occu(s)})$, where Occu(s) is the number of occurrences of the sense s. The second set consists of features (f) that co-occur with multiple senses and are ordered according to their log-likelihood ratio: $\log(\frac{Occu(s,f)}{Occu(f)-Occu(s,f)})$. Given a novel instance, the first set is applied first; if the sense cannot be determined by the first set, the second set is then applied; and the default test returns the majority sense.

Observing the existence of instances with rare senses deteriorates Naïve Bayesian classifiers, our **mixed supervised learning algorithm**, which contains a Naïve Bayesian classifier and an instance-based classifier, was implemented. The algorithm can be stated as follows:

Mixed Supervised Learning Algorithm

- Split the training set to two parts, I and II, where part I contains instances with majority senses, and part II contains instances with rare senses (see the footnote at Page 17 for definitions of rare senses)
- Build a Naïve Bayes classifier trained on part I and an instance-based classifier trained on part II

- For a novel instance, if the instance-based classifier predicates its sense with a relatively high similarity, return the predicated sense; else return the predicate sense of the Naïve Bayes classifier

Note that if there are no rare senses in the training set, our mixed supervised learning algorithm is the same as Naïve Bayes learning.

6.1.4. Methods

For each ambiguous abbreviation *AW* in Set A, we derived 70 WSD classifiers: 10 were represented using a pair (**ml**, $\mathbf{fp}_1$), and 60 were represented by a tuple (**ml**, $\mathbf{fp}_2$, **ws**). The variable **ml** is a supervised learning algorithm with five choices: Naïve Bayes learning, traditional decision list learning, our implementation of decision list learning, instance-based learning, and our mixed supervised learning. $\mathbf{fp}_1$ and $\mathbf{fp}_2$ are feature presentation variables, where $\mathbf{fp}_1$ has two values "d" and "f", and $\mathbf{fp}_2$ has four values "a", "b", "c", and "e" (refer to Table 6 for these feature representations). The variable **ws** is the window size with three values (3, 5, 10)[18].

Since instances in the gold standard set of *AW* come from abstracts, there are multiple occurrences of *AW* in some abstracts. Based on the fact that all occurrences of an abbreviation in an abstract generally have the same sense, we assigned all occurrences of *AW* the gold standard sense. Features were extracted for all occurrences in the gold standard set of *AW,* and all measures were computed for occurrences[19].

[18] We could test every possible window size. We chose these three values to see the preference of window sizes.

[19] For abbreviations with over 6,000 instances in the gold standard set, we randomly chose 5,000 instances for the experiment because it took more than 3 hours to finish the 10-fold cross-validation process for those words.

For each ambiguous term *W* in Set B, we derived 86 WSD classifiers: 6 were represented using a pair (**ml**, $\mathbf{fp_1}$), and 80 were represented by a tuple (**ml**, $\mathbf{fp_2}$, **ws**). The variable **ml** is a supervised learning algorithm with four choices: Naïve Bayes learning, traditional decision list learning, our implementation of decision list learning, and instance-based learning. Note that we excluded our mixed supervised learning here since there were no rare senses in the gold standard sets for terms in Set B and our mixed supervised was the same as Naïve Bayes learning. $\mathbf{fp_1}$ and $\mathbf{fp_2}$ are feature presentation variables, where $\mathbf{fp_1}$ has two values "d" and "f", and $\mathbf{fp_2}$ has four values "a", "b", "c", and "e" (refer to Table 6 for these feature representation values). The variable **ws** is a window size with five values (2, 3, 4, 5, 10). The instances in the gold standard set of *W* were sentences, features were extracted for each sentence and measures were reported using sentences.

We applied the 10-fold cross-validation method to measure the performance. Measures were averaged over the results of the 10 folds.

6.1.5. Results

Observing instance-based classifiers required a long time to execute and preliminary analysis showing that instance-based classifiers had poor performance, we aborted all Instance-Based classifiers.

The overall performance of different classifiers for sets A and B is listed in Table 7 and Table 8, respectively. Classifiers with the best overall performance for Set A were classifiers using feature representation "f" and three supervised learning algorithms: Naïve Bayes learning, our implementation of decision list learning, and our mixed supervised learning. The performance of those classifiers was significantly better than

that of other classifiers and achieved an overall precision of over 99%. Classifiers with the best overall performance for Set B were decision list classifiers disregarding feature representations and window sizes. The classifier with the worst overall performance for Set A used feature representation "d" and Naïve Bayes learning; the precision of the classifier was about 82%. The classifier with the worst overall performance for Set B also used Naïve Bayes learning but with feature representation "a" and a window size of 10.

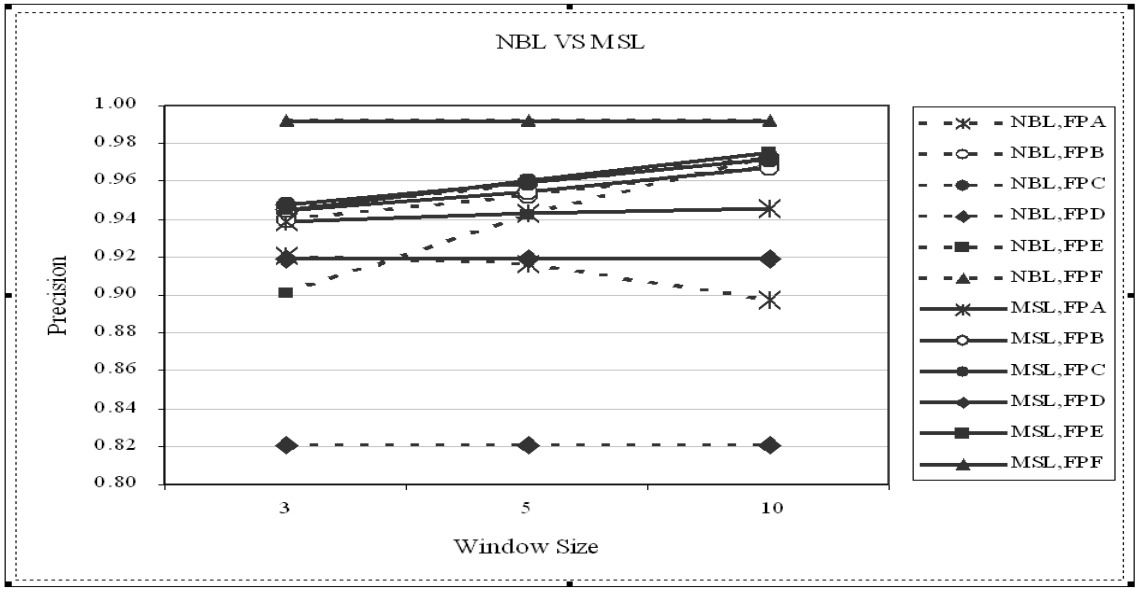
The comparison of two decision list learning algorithms and the comparison of Naïve Bayesian algorithm with our mixed supervised learning algorithm are shown in Figure 7 for Set A. Figure 7 also shows the relation of the overall performance of classifiers with different window sizes. Figure 8 shows the relation of different combinations of supervised learning algorithms and feature representations with window sizes for Set B. Note that feature representation "a" is denoted as FPA, and so forth for representations "b", "c", "d", "e", and "f". From figures 7 and 8, we found that Naïve Bayes learning was unstable and varied dramatically for different feature representations. For a fixed window size **ws** and a fixed feature representation option **fp**, the performance of our implementation of decision list classifiers for Set A was significantly better than traditional decision list classifiers except when the value of **fp** was "a"; the performance of mixed supervised learning classifiers for Set A was generally superior than that of Naïve Bayes classifiers; the performance of both implementations of decision list classifiers for Set B was much better than Naïve Bayes classifiers while the performance of our implementation of decision list learning was slightly but not significantly worse than that of traditional decision list learning for words in Set B.

FP	WS	Machine Learning Algorithm P (0.95 CI)			
		TDLL	MYDLL	NBL	MSL
a	3	94.2 (94.1-94.3)	94.1 (94.0-94.2)	92.0 (91.9-92.2)	93.9 (93.8-94.0)
	5	94.3 (94.2-94.4)	94.1 (94.0-94.2)	91.6 (91.5-91.7)	94.3 (94.2-94.4)
	10	94.2 (94.1-94.3)	94.1 (94.0-94.2)	89.7 (89.6-89.8)	94.6 (94.5-94.7)
b	3	94.2 (94.1-94.3)	94.3 (94.2-94.4)	93.9 (93.8-94.0)	94.5 (94.4-94.5)
	5	94.7 (94.6-94.8)	94.9 (94.8-95.0)	95.3 (95.2-95.3)	95.5 (95.4-95.6)
	10	95.5 (95.4-95.5)	95.9 (95.8-96.0)	96.8 (96.7-96.9)	96.8 (96.7-96.8)
c	3	94.3 (94.2-94.4)	94.4 (94.3-94.5)	94.5 (94.4-94.6)	94.8 (94.7-94.9)
	5	95.0 (94.9-95.1)	95.2 (95.1-95.3)	95.9 (95.8-95.9)	95.9 (95.8-96.0)
	10	95.9 (95.8-96.0)	96.4 (96.3-96.4)	97.3 (97.2-97.3)	97.2 (97.1-97.3)
d	NA	94.6 (94.5-94.7)	94.7 (94.7-94.8)	82.1 (81.9-82.2)	91.9 (91.8-92.0)
e	3	94.8 (94.7-94.9)	95.1 (95.0-95.2)	90.1 (89.9-90.2)	94.5 (94.4-94.6)
	5	95.4 (95.3-95.5)	95.8 (95.7-95.9)	94.2 (94.1-94.3)	96.0 (95.9-96.1)
	10	96.2 (96.1-96.3)	96.7 (96.7-96.8)	97.3 (97.3-97.4)	97.5 (97.4-97.5)
f	NA	98.5 (98.5-98.6)	99.2 (99.1-99.2)	99.2 (99.2-99.3)	99.1 (99.1-99.2)

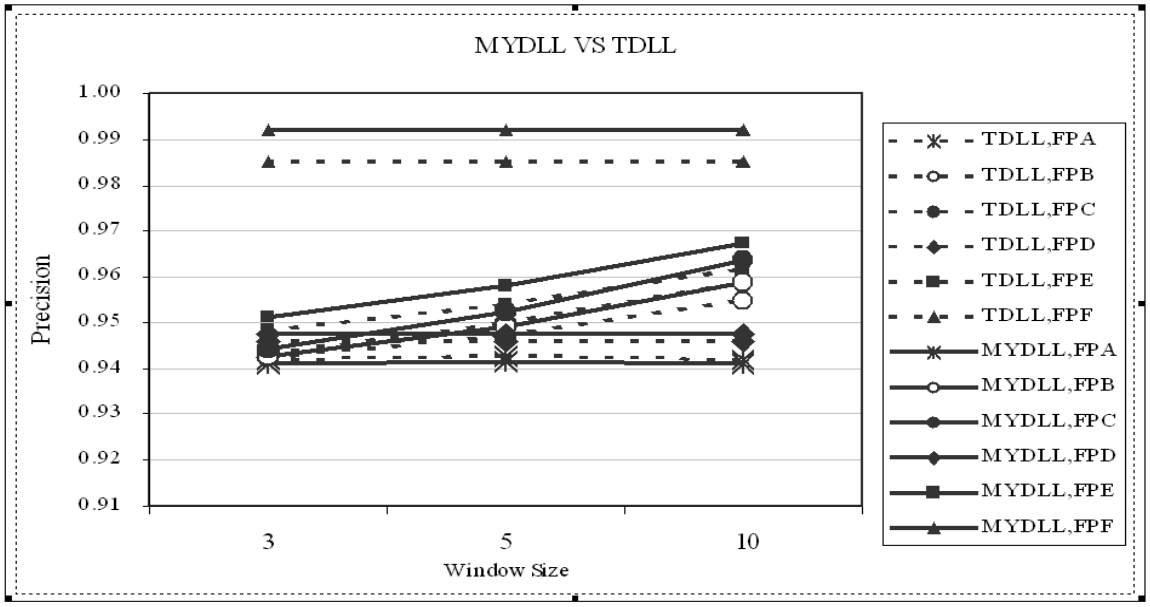
Table 7. The overall precision of different classifiers for abbreviations in Set A (i.e., 35 frequent abbreviations, see Section 5.3). The machine learning algorithm has four choices traditional decision list algorithm (DLL), our implementation of decision list algorithm (MYDLL), Naïve Bayesian algorithm (NBL), and mixed supervised learning algorithm (MSL); the feature presentation (FP) has six options: “a”, “b”, “c”, “d”, “e”, and “f”, where representations “a”, “b”, “c”, and “e” have three different window sizes (WS) 3, 5, and 10.

FP	Machine Learning	Overall Precision (%)				
		Window Size				
		2	3	4	5	10
a	TDLL	87.9	87.4	87.2	86.5	86.0
	MYDLL	87.7	86.6	85.7	84.6	82.1
	NBL	66.2	56.4	49.9	45.3	34.5
b	TDLL	88.1	87.7	87.5	87.3	87.7
	MYDLL	88.2	87.4	87.3	86.6	85.1
	NBL	69.1	64.3	62.9	61.5	63.5
c	TDLL	87.6	87.5	87.7	87.9	88.5
	MYDLL	87.5	87.6	87.7	87.1	87.4
	NBL	71.7	68.8	69.3	69.1	74.2
d	TDLL	88.4	88.4	88.4	88.4	88.4
	MYDLL	88.4	88.4	88.4	88.4	88.4
	NBL	47.7	47.7	47.7	47.7	47.7
e	TDLL	87.8	87.3	87.4	87.9	88.8
	MYDLL	87.7	86.8	87.0	87.0	87.8
	NBL	53.8	54.3	56.3	57.0	65.3
f	TDLL	89.4	89.4	89.4	89.4	89.4
	MYDLL	87.6	87.6	87.6	87.6	87.6
	NBL	73.5	73.5	73.5	73.5	73.5

Table 8.The overall performance of different classifiers for words in Set B (i.e., general biomedical terms used in the WSD project of NLM, see Section 5.3). The machine learning algorithm has three choices TDLL, MYDLL, and NBL (refer to Table 7 for definitions of TDLL, MYDLL, and NBL); the feature presentation (FP) has six options: “a”, “b”, “c”, “d”, “e”, and “f”, where “a”, “b”, “c”, and “e” have five different window sizes 2, 3, 4, 5, and 10.



(I)



(II)

Figure 7. I). Comparison of NBL with MSL and their relation to window sizes for Set A; II). Comparison of TDLL with MYDLL and their relation to window sizes for Set A.

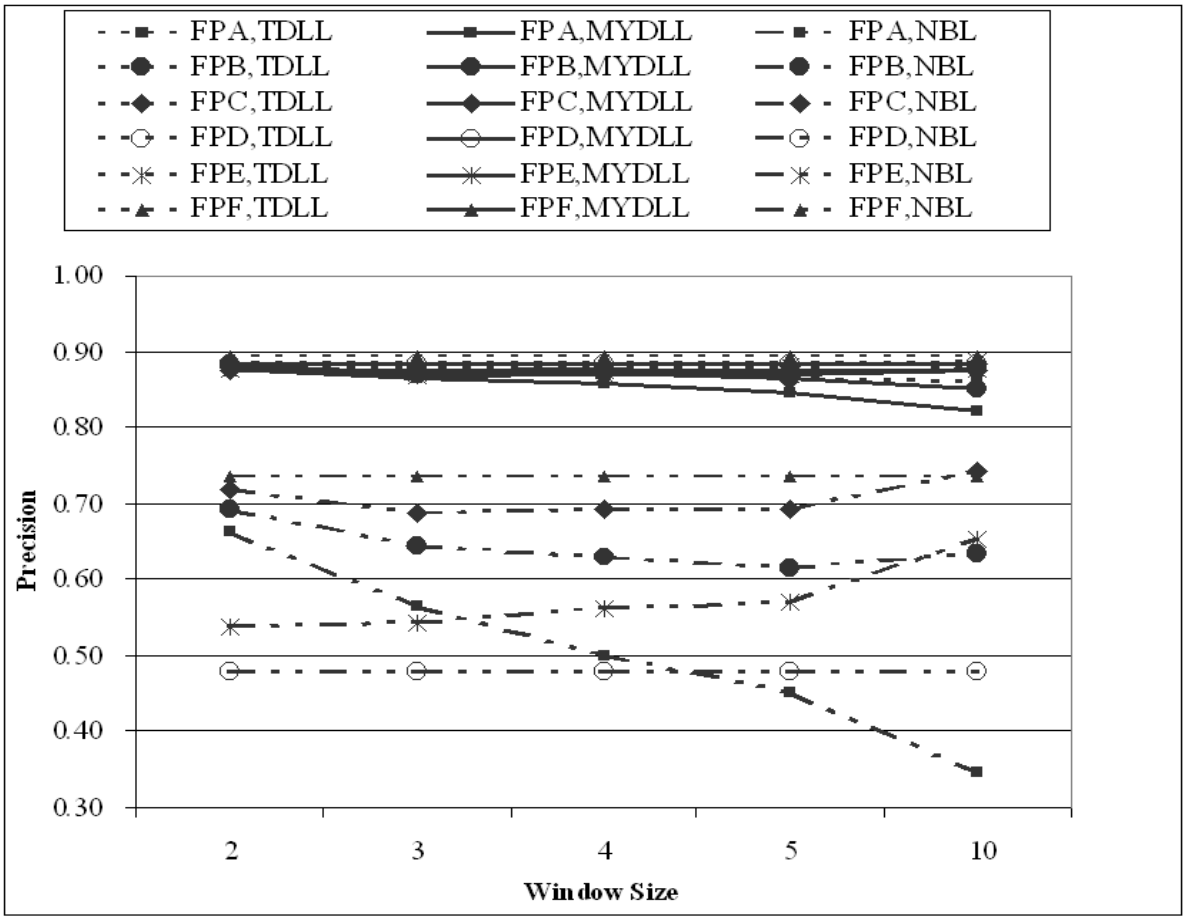


Figure 8. Relations between classifiers and window sizes for Set B.

The parameters of the best classifier for each term from sets A and B are listed in Table 9 and Table 10, respectively. Tables 9 and 10 also list the corresponding precision of the classifier and the precision of the best overall classifier. The best classifiers for almost all abbreviations in Set A were using feature presentation "f" (i.e., all other words appearing in the context and collocations) and all supervised learning algorithms except traditional decision list learning, with a precision of almost 100%. However, there is no regularity of the best classifier for terms in Set B.

AW	Best Classifier				BOC (%)
	FP	WS	ML	Precision	Precision
ACE	f	NA	NBL	100.0	100.0
ANA	f	NA	{NBL, MSL, MYDLL	100.0	100.0
APC	f	NA	NBL	99.9	99.9
ASP	f	NA	{NBL, MSL}	99.3	99.3
BPD	f	NA	{NBL, MSL}	99.9	99.9
BSA	f	NA	{NBL, MSL}	99.9	99.9
CAD	f	NA	NBL	100.0	100.0
CAT	a	{5, 10}	{DLL,MYDLL}	98.8	97.7
CML	f	NA	{NBL, MSL}	99.7	99.7
CMV	f	NA	MYDLL	100.0	100.0
CPI	f	NA	{NBL, MSL, MYDLL}	100.0	100.0
CSF	f	NA	NBL	99.9	99.9
CVA	*	*	*	100.0	100.0
CVP	f	NA	{NBL, MSL, MYDLL}	99.9	99.9
DIP	f	NA	*	100.0	100.0
DOB	a	NA	NA	NA	NA
DVT	f	NA	NBL	99.9	99.9
EMG	f	NA	{NBL, MSL}	87.7	87.7
FDP	f	NA	{NBL, MYDLL}	99.9	99.9
HSV	f	NA	{NBL, MSL}	100.0	100.0
IBD	*	*	*	100.0	100.0
LAM	f	NA	{NBL, MSL}	99.4	99.4
LDH	f	NA	{NBL, MYDLL}	100.0	100.0
MAC	f	NA	{NBL, MSL, MYDLL}	99.9	99.9
MAS	f	NA	*	100.0	100.0
MCP	f	NA	{NBL, MSL, MYDLL}	99.9	99.9
PCA	f	NA	NBL	99.9	99.9
PCP	f	NA	MYDLL	99.8	99.8
PEG	f	NA	{NBL, MSL}	99.6	99.6
PSA	f	NA	MYDLL	100.0	100.0
PVC	f	NA	{NBL, MSL}	99.7	99.7
RSV	f	NA	MYDLL	99.7	99.7
SLE	d	NA	MYDLL	99.3	99.3
TPN	f	NA	MYDLL	100.0	100.0
VCR	f	NA	*	99.9	99.9

Table 9. The parameters of the best classifiers and their precisions as well as the precision of the best overall classifiers (BOC) for each word in Set A. * here means any value for that variable. Refer to Table 7 for notations.

WORD	Best Classifier				BOC(%)
	FP	WS	ML	Precision	Precision
ASSOCIATION	*	*	*	100.0	100.0
COLD	c	4	MYDLL	93.7	91.1
CULTURE	c	5	MYDLL	96.2	88.5
DEGREE	d	NA	TDLL	98.3	96.6
DEPRESSION	a	3	TDLL	94.6	85.7
DISCHARGE	{e,f}	10	NBL	90.4	76.7
ENERGY	*	{2,4}	{TDLL, MYDLL}	100.0	99.1
EXTRACTION	d	NA	MYDLL	89.3	67.7
FAT	{a,b,f}	4	{MYDLL, TDLL}	87.9	87.9
FIT	{c,e}	{4,5}	MYDLL	94.3	90.7
FLUID	*	*	*	100.0	100.0
FREQUENCY	e	2	TDLL	96.0	93.1
GANGLION	a	2	MYDLL	98.1	91.5
GLUCOSE	e	10	TDLL	94.3	92.7
GROWTH	f	NA	{TDLL, NBL}	75.5	75.5
IMPLANTATION	f	NA	TDLL	90.5	90.5
INHIBITION	c	3	MYDLL	98.1	98.1
JAPANESE	c	3	MYDLL	84.5	81.9
LEAD	d	NA	MYDLL	88.3	85.9
MAN	a	2	MYDLL	91.0	79.2
MOLE	{a,f}	10	TDLL	100.0	100.0
NUTRITION	e	10	NBL	76.7	73.3
PATHOLOGY	epc	10	MYDLL	89.1	83.7
PRESSURE	a	4	TDLL	98.2	96.4
REDUCTION	e	10	MYDLL	92.0	90.9
REPAIR	f	NA	TDLL	80.2	80.2
RESISTANCE	f	NA	MYDLL	97.1	96.1
SCALE	{e,f}	10	NBL	91.9	84.8
SECRETION	{a,b,c}	3	{TDLL, MYDLL}	100.0	98.9
SEX	b	5	MYDLL	94.0	90.1
SINGLE	b	4	TDLL	100.0	98.9
STRAINS	f	NA	MYDLL	95.4	92.3
SURGERY	{c,f}	4	TDLL	99.0	99.0
TRANSIENT	*	*	*	100.0	100.0
TRANSPORT	a	10	{TDLL, MYDLL}	100.0	99.0
ULTRASOUND	b	2	MYDLL	88.4	83.3
WEIGHT	{c,e}	{5,10}	{TDLL, MYDLL}	83.9	76.8
WHITE	b	5	TDLL	81.3	70.5

Table 10. The parameters of the best classifiers and their precisions as well as the precisions of the best overall classifiers (BOC) for each word in Set B. Refer to tables 7, 8 and 9 or notations.

6.1.6. Discussion

The instance-based learning took a very long time to execute. One possible reason may be that in the current study, there were over thousands of instances in the training and the test sets for each abbreviation. The other possible reason is that our implementation may not be efficient even though it ran well for a small size single word disambiguation task (about 600 instances in the gold standard set) where the running time for the process was about 3 hours.

We found that the performance of WSD classifiers was related to the related-ness of senses. For terms with closely related senses, supervised WSD classifiers had a lower precision compared to terms with unrelated senses. For example, there were four senses of *EMG*: EMG_1 (i.e., *exomphalos macroglossia gigantism*), EMG_2 (i.e. *electromyography*), EMG_3 (i.e. *electromyographs*) and EMG_4 (i.e. *eletromyogram*); three of them, i.e, EMG_2, EMG_3 and EMG_4, were closely related and the best supervised WSD classifier for EMG had a precision of 87.7% compared to a precision of almost 100% for other abbreviations. In addition, we believe the existence of closely related senses is one of the reasons that WSD classifiers for abbreviations in Set A had better performance than WSD classifiers for general biomedical terms in Set B.

Furthermore, the performance of WSD classifiers was related to the number of instances in the sense-tagged corpora. For example, the gold standard set for each abbreviation in Set A usually contained over thousands of instances while there were only dozens of instances for each term in Set B. Abbreviations in Set A were disambiguated with a higher precision compared to terms in Set B. The power of supervised machine learning algorithms is that they can gather disambiguation knowledge from a large number of

instances in the training set. With only dozens of instances for each term in Set B, supervised machine learning techniques seem to be inappropriate.

6.1.7. Conclusions

We conducted an experiment that compared different feature representations and different machine learning algorithms. Our results demonstrated that supervised WSD classifiers have a reliable performance when there exist a large number of sense-tagged instances. Feature representations including collocations and neighboring words are appropriate representations for the context. For terms with domain-specific senses, a large window size should be used. For general English terms, a small window size of 2 to 5 should be used. Our mixed supervised learning was stable and generally had better performance than Naïve Bayes learning for abbreviations, and our implementation of decision list learning had a better performance for traditional decision list learning.

6.2. Noise Tolerance of Supervised Learning Algorithm

From the previous experiment, we found that our implementation of decision list learning was superior than traditional decision list learning for abbreviations in Set A, and that our mixed supervised learning has a better performance than Naïve Bayes learning. In this section, we discuss the performance of different supervised learning when noise is presented in the training set.

6.2.1. Methods

For each abbreviation in Set A, we randomly split the gold standard set into a training set and a test set with a ratio 9:1. Nine training sets were derived from the training set, where each had a level of noise from nine levels, i.e., 0%, 5%, 10%, 15%, 20%, 25%, 30%, 35%

and 40%. For example, when the noise level is 5%, we substituted the gold standard sense for 5% of the instances in the training set to a randomly chosen different sense. Nine WSD classifiers were built using each of the derived training sets. We used a feature representation that was similar to feature representation "f", which contained features in feature representation "d" for all occurrences of *AW* in an instance and all words except *AW* in the instance. We compared four different supervised learning algorithms: tradition decision list learning, our implementation of decision list learning, Naïve Bayes learning, and our mixed supervised learning. The performance was measured for abstracts, i.e., we reported measures using abstracts instead of using occurrences as in Section 6.1.

We report on the overall performance of each supervised learning algorithm for different noise level. The measures were averaged over five separate runs.

6.2.2. Results

Detailed information about the best classifier for each combination of abbreviation and noise is presented in Appendix F. The overall performance for each supervised learning algorithm is shown in Figure 9. The tolerance of noise was different among the different supervised learning algorithms. Naïve Bayes learning had a low precision when there was noise in the training set for abbreviations with a skewed sense distribution or having rare senses. Our implementation of decision list learning had a lower precision compared to traditional decision list learning. Traditional decision list learning was robust and had the best performance for abbreviations with a skewed sense distribution. Our mixed supervised learning had the best performance for abbreviations with a balanced sense

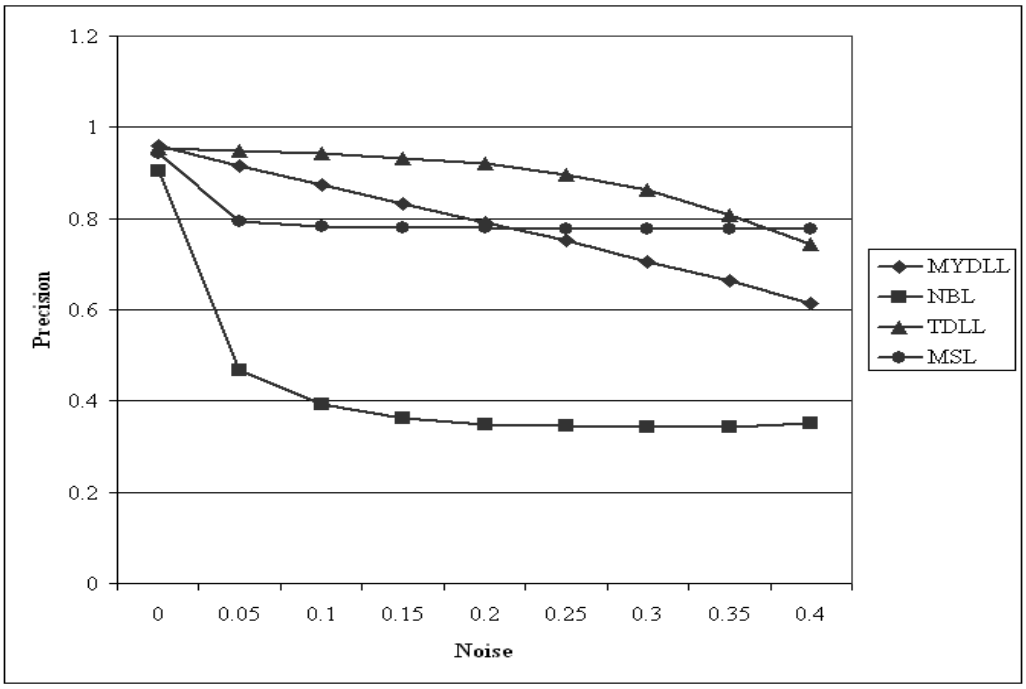


Figure 9. Comparison of the overall performance of different supervised learning algorithms in relation to different noise levels.

distribution. Note that the drop at 0.05 of our mixed supervised learning in Figure 9 was caused by two abbreviations, *ACE* and *CAD*, which has only one majority sense.

6.2.3. Discussion

In our study, we replaced the gold standard sense with other senses using an equal probability that may not reflect the true distribution of noise in the training set for real world applications (e.g., rare senses in the gold standard set most likely became non-rare senses in our study, which might not happen in the real world). Our mixed supervised learning algorithm was very robust when the sense distribution is balanced. The performance of our mixed supervised classifiers trained on the training set with 40% of noise was not significantly different from that of our mixed supervised classifiers when

trained on the gold standard training set for abbreviations with a balanced sense distribution. The performance of traditional decision list classifiers trained on the training set with 10% of noise was not significantly different from that of decision list classifiers trained on the gold standard training set for abbreviations with a skewed sense distribution.

6.3. Construction of WSD Classifiers for Abbreviations

We hypothesized that preliminary sense-tagged corpora (STC') for abbreviations, which are derived automatically, are comprehensive and can be used to derive WSD classifiers directly without two optional components (i.e. without the clustering analysis component and the expert annotation component). We answered the hypothesis through a set of experiments. In experiment I, we built WSD classifiers for abbreviations in Set A using sense-tagged corpora that were derived from unambiguous synonyms (STC_1). The performance of those WSD classifiers was evaluated using the gold standard set. In experiment II, we evaluated the quality of sense-tagged corpora derived using conceptual relatives in the context. In experiment III, we addressed the performance of WSD classifiers trained on sense-tagged corpora that were derived automatically. In addition, we studied the transferability of the constructed WSD classifiers, which used MEDLINE abstracts as training instances, by applying them on the disambiguation task of abbreviations in medical reports.

6.3.1. Experiment I

6.3.1.1. Methods

For each abbreviation *AW* in Set A, a raw corpus, $RC_1(AW)$, where each instance in the corpus contained an unambiguous synonym of *AW*, was derived with the following restriction: each synonym can only be used to extract 5,000 abstracts for abbreviations. After substituting all occurrences of the corresponding synonym in each abstract of *AW* with *AW* and tagging *AW* with the associated sense of that synonym, we derived a sense-tagged corpus, $STC_1(AW)$ (see Section 4.1.4). We excluded all abstracts that were in the gold standard set of *AW* from $STC_1(AW)$ and the result was the training set of *AW*.

For each abbreviation *AW*, a WSD classifier for *AW* was then trained on the training set and tested on the gold standard set of *AW*. We compared four different supervised learning algorithms, i.e., Naïve Bayes learning, traditional decision list learning, our implementation of decision list learning, and our mixed supervised learning.

6.3.1.2. Results

The information about the training set for each abbreviation in Set A is listed in the second column of Table 11: NS is the number of presented senses, TR is the number of instances, MAJ is the majority sense, and the PMAJ is the percentage of the majority sense (for the detail information for each sense, see Appendix C). The distribution of senses in the training set was different from that in the test set for most words. For example, the majority sense in the training set for *ACE* was ACE_1 (i.e., *acetylcholinesterase*); while in the gold standard set, the majority sense was ACE_2 (i.e., *angiotensin converting enzyme*). The sizes of the training sets were quite different among

abbreviations: some (i.e. *APC, BPD, BSA, HSV, IBD, RSV, TPN*) had over 300 instances for each sense; while others had at least one sense with less than 300 instances. Not all senses with instances in the gold standard set had instances in the training sets. For example, there were eight instances in the gold standard set of *MCP* associated with the sense MCP_2; but there were no instances in the training set associated with the sense MCP_2.

The precision for each supervised learning algorithm is listed in the third column of Table 11; the precision of the classifiers that were significantly better are flagged with stars. The performance of WSD classifiers varied for different abbreviations. For example, all WSD classifiers for *HSV* achieved a precision of over 99%, while all WSD classifiers for *DIP* achieved a precision of 27.7%. The best WSD classifiers for each abbreviation were also different from each other. Among 35 abbreviations in Set B, 13 (e.g., *APC*, *BPD*, etc.) had a precision of over 90% for all the WSD classifiers. However, there were 6 abbreviations (i.e., *DIP*, *DOB*, *EMG*, *LAM*, *MAC*, and *PCP*) for which all WSD classifiers had a precision of less than 90%.

The supervised learning algorithm with the best overall performance was mixed supervised learning algorithm with an overall precision of 90.4%. The overall precision of our implementation of decision list learning and the mixed supervised learning was significantly better than traditional decision list learning and Naïve Bayes learning with a 95% confidence interval.

AW	TRAINING SET				PRECISION (%)			
	NS	TR	MAJ	PMAJ(%)	TDLL	MYDLL	NBL	MSL
ACE	4	8,583	ACE_1	62.8	94.5	93.3	79.9	97.5*
ANA	3	254	ANA_2	57.1	98.8*	97.8*	93.3	93.3
APC	5	6,682	APC_2	33.8	98.1*	96.1	98.9*	98.9*
ASP	6	13,824	ASP_6	34.9	84.4	77.3	92.2	92.9
BPD	3	2,302	BPD_1	41.0	97.6	97.6	99.4*	99.4*
BSA	2	66,94	BSA_2	51.5	96.6	96.1	97.8*	97.8*
CAD	2	12,182	CAD_1	99.7	99.1	99.5*	98.8	99.1
CAT	4	19,121	CAT_2	85.7	94.4	97.2	100	97.2
CML	2	3,750	CML_1	99.5	94.9*	95.6*	93.7*	94.9*
CMV	3	190	CMV_4	76.3	44.3	90.0*	87.3	87.3
CPI	3	143	CPI_2	52.4	94.4	97.2	95.8	95.8
CSF	3	7,133	CSF_1	56.4	97.4*	95.6	89.0	89.0
CVA	2	10,489	CVA_1	99.6	100.0*	100.0*	83.2	100*
CVP	3	14	CVP_3	85.7	99.3*	44.3	13.1	99.0*
DIP	1	69	DIP_2	100.0	27.7	27.7	27.7	27.7
DOB	1	535	DOB_3	100.0	50.0	50.0	50.0	50.0
DVT	2	1,745	DVT_1	97.1	99.1*	99.1*	83.9	83.9
EMG	3	3,248	EMG_2	96.8	21.4	21.9	52.3	52.3
FDP	3	676	FDP_3	67.2	97.2*	97.4*	83.5	97.4*
HSV	2	6,339	HSV_1	91.4	99.5	99.9	99.7	99.7
IBD	2	5,386	IBD_1	72.7	97.2*	93.0	81.2	81.2
LAM	4	4,702	LAM_2	93.0	9.8	65.6	82.5*	82.5*
LDH	2	91	LDH_1	96.7	100.0*	99.2*	75.9	100*
MAC	7	9,975	MAC_2	85.3	37.8	66.6	89.1*	86.7*
MAS	3	564	MAS_2	74.3	99.1	99.1	100.0	100
MCP	3	2,650	MCP_4	92.5	62.0	92.0	90.5	55.5
PCA	7	3,659	PCA_8	46.6	73.0	85.3	92.5*	80.6
PCP	5	2,462	PCP_5	74.6	49.8	77.9	84.3	51.4
PEG	2	4,517	PEG_1	92.8	77.1	92.9*	100.0*	100*
PSA	3	1,596	PSA_1	62.0	99.5	97.8	94.7	99.8
PVC	3	7,109	PVC_2	50.1	97.7*	97.2*	47.3	97.7*
RSV	2	2,672	RSV_2	55.5	95.6	95.2	96.1	96.1
SLE	2	2,344	SLE_2	99.4	97.7*	99.7*	85.8	97.7*
TPN	2	7,601	TPN_1	74.3	85.3	90.8	98.7*	98.7*
VCR	3	4,598	VCR_1	99.7	99.4*	99.1*	95.8	99.4*
Total	NA				84.8	89.7	86.5	90.4

Table 11. The evaluation results of STC_1 using Set A. NS (TR) - the number of presented senses (instances) , MAJ (PMAJ) - the majority sense (the percentage of the occurrences). Refer to Table 7 for other notations.

6.3.1.2 Discussion

We analyzed the causes of the low precision of WSD classifiers, and there were several causes: too few sense-tagged instances and relatedness among different senses.

The cause for the low precision of WSD classifiers was that there were not enough sense-tagged instances. For example, 72.7% of instances in the gold standard set of *DIP* held a sense that was not presented in the training set, and 50% of instances in the gold standard set of *DOB* held a sense that was not present in the training set. There were only 14 instances in the training set of *PCP* having the sense PCP_4, while there were 1,071 instances in the gold standard set of *PCP* with the sense PCP_4.

The low precision of some WSD classifiers was caused by the existence of closely related senses for the corresponding abbreviations. For example, all WSD classifiers for *EMG* had a precision of less than 55%.

We found that if there were enough instances (e.g. over 300 instances) for each sense, almost all WSD classifiers achieved a high precision regardless of the level of ambiguity and sense distributions. For example, the abbreviation *APC* had five senses, each with over 500 instances (refer to Appendix C), and all WSD classifiers for *APC* achieved a precision of over 96%. The abbreviation *BPD* had three senses with over 400 instances for each, and all WSD classifiers for *BPD* achieved a precision of over 97%. However, the results were different for abbreviations with closely related senses. For example, two senses of *IBD* were closely related (i.e., IBD_1 for *inflammatory bowel disease* and IBD_2 for *irritable bowel syndrome*) and the Naïve Bayes classifier for *IBD* had a precision of

81.2% even though there were 3,916 instances with the sense IBD_1 and 1,470 instances had the sense IBD_2.

We observed for almost all abbreviations with a skewed sense distribution (i.e., the majority sense occurred more than 90%), such as *CAD*, *CML*, *SLE*, and *VCR*, or abbreviations with two closely related senses, such as *IBD* and *DVT*, decision list learning achieved the best performance. Naïve Bayes classifiers achieved the best performance for words with a balanced sense distribution, such as *BPD*, *APC*, etc. However, they were not robust when there were rare senses (i.e., senses that occurred less than 1%). For example, only 3 instances in the training set of *ACE* had rare senses, but the performance of Naïve Bayes classifier was much worse than other learning algorithms. Our mixed supervised learning algorithm that separated rare senses from other senses was robust, and achieved the best performance for words with a training set that had a balanced sense distribution for non-rare senses, such as *ACE* and *PVC* etc.

6.3.2. Experiment II

6.3.2.1. Methods

For each abbreviation *AW* in Set A, a raw corpus which consisted of MEDLINE abstracts containing *AW*, was extracted from MEDLINE. For each abstract, conceptual relatives were identified using CRMap (refer to Section 4.1.4.3). Since not every abstract had identified conceptual relatives, we measured the performance of each individual relation type (or source) using the following measures: coverage (i.e., the number of abstracts that had occurrences of conceptual relatives with that type (or source)), and precision (i.e., the

percentage of the number of occurrences of conceptual relatives with that type (or source) that correctly identified the sense).

We derived a sense-tagged corpus, $STC_2(AW)$, using the majority vote sense of those identified relatives; if there was a tie, we randomly chose one of the tied senses. We evaluated the quality of the corpus using the gold standard set of *AW* with two measures: recall, i.e. the ratio of the number of abstracts with correctly identified sense to the total number of abstracts in the evaluation set, and precision, the ratio of the number of abstracts with correctly identified sense to the number of abstracts that were assigned senses.

We also compared the quality of the sense-tagged corpus using two different mapping programs, i.e., CRMap and MetaMap, for several abbreviations.

6.3.2.2. Results

We extracted 155,723 abstracts from MEDLINE. There were 85,554 abstracts with conceptual relatives identified using CRMap (see Appendix C for detailed statistics about each sense).

The overall performance of different relation types (or sources) is shown in Table 12 (note only sources with over 1% of the coverage were included in the Table) when evaluated using the gold standard set. Among 19 different sources with a coverage of over 1%, MTH (i.e., relations created during the construction of the META) had the highest coverage (i.e., 27.4%) and CCPSS99 (i.e., relations imported from the Canonical Clinical Problem Statement System, 1999 version) had the highest precision (98.7%). The conceptual relatives with relations from 2 sources (i.e., WHO97 (i.e., WHO Adverse

SOURCE	COV (%)	PRE (%)
MTH	27.4	91.8
MSH2001	24.7	91.5
CSP2000	21.8	92.2
CST95	11.1	85.4
CCPSS99	10.4	98.7
RCD99	10.0	94.0
SNMI98	10.0	89.9
AOD99	8.3	94.3
WHO97	6.8	65.3
BI98	4.9	97.7
META	4.3	93.5
PSY97	2.4	93.5
PDQ2000	2.0	74.8
LNC10o	1.6	95.7
ICD10AM	1.4	99.6
ICD2001	1.3	99.4
ICPC2P	1.1	99.3
CCS99	1.1	99.0
SNM2	1.0	99.7

TYPE	COV (%)	PRE (%)
Other	34.4	94.6
Sibling	21.4	82.6
Child	15.4	91.3
Narrower	14.2	88.0
Parent	9.0	91.6
Broader	8.9	96.3
Synonymy	4.3	99.7
Like	0.5	94.6

Table 12. Comparison results among different sources and types. COV stands for coverage and PRE stands for precision[20].

Drug Reaction Terminology, 1997 version) and PDQ2000 (i.e., Physician Data Query Online System, 2000 version)) had a precision of lower than 80%; conceptual relatives from 15 had a precision of over 90%. Among 8 different relation types, the conceptual relatives with a relation type, Other, achieved the best coverage (34.4%); while the synonymy (i.e., the relation assigned to terms with the same conceptual identifier) relation achieved the best precision (99.7%). The conceptual relatives from 6 types (i.e.,

[20] Refer to http://www.nlm.nih.gov/research/umls/2002AB_Addendum.html for descriptions of Sources

Other, Child, Parent, Broader, Synonymy, Like) had a precision of over 90%. The conceptual relatives with the Sibling type had the lowest precision (82.6%).

The average recall of the sense tagged corpus was 48.0% when evaluated on the gold standard set, and the average precision was 92.5%. Table 12 lists the detailed information about STC_2 and measures for each abbreviation, where STC_2-R is the recall and STC_2-P is the precision. For example, the number of abstracts in STC_2(*ANA*) is 1,101, and about 73.3% of them have been assigned correct senses; all sense assignments of *ANA* were correct based on concept relatives. From Table 13, we can see that the measures differed widely among abbreviations. For example, the STC_2 for 27 out of 35 abbreviations had a precision of over 94% (e.g. *ACE*, *CAD*, etc), while there were 4 abbreviations (i.e. *ASP*, *DVT*, *EMG*, and *MAC*), where STC_2 had a precision of lower than 80%. The result of the comparison between CRMap and MetaMap is listed in Table 14. CRMap was significantly better than MetaMap with respect to the quality of STC_2, except for *BSA*, which had a lower recall (9% compared to 29.9%).

6.3.2.3. Discussion

In this dissertation, we assigned the sense of an abbreviation in an abstract as the majority vote sense of conceptual relatives in the abstract. A more sophisticated sense assignment mechanism can be developed by assigning weights to different sources, types, or combinations of these two.

We analyzed the causes of low precision for STC_2, and there were two causes: relatedness among different senses and the existence of poor conceptual relatives.

AW	STC_2	STC_2-R (%)	STC_2-P (%)	AW	STC_2	STC_2-R (%)	STC_2-P (%)
ACE	7,460	76.7	97.9	FDP	727	55.0	100.0
ANA	1,101	73.3	100.0	HSV	6,494	38.9	99.9
APC	4,158	68.8	84.3	IBD	1,191	80.7	96.2
ASP	305	63.8	74.4	LAM	146	30.6	87.5
BPD	494	39.5	97.5	LDH	4,032	48.3	100.0
BSA	1,408	9.0	89.9	MAC	1,382	58.5	78.3
CAD	3,501	85.5	99.9	MAS	110	60.7	98.6
CAT	1,584	41.7	100.0	MCP	2,004	72.2	98.2
CML	3,169	61.5	99.0	PCA	676	22.7	94.4
CMV	4,570	63.1	99.4	PCP	1,736	50.6	94.5
CPI	32	22.2	100.0	PEG	1,220	34.3	100.0
CSF	21,497	38.4	88.6	PSA	1,396	28.0	98.5
CVA	402	76.1	100.0	PVC	438	25.4	94.2
CVP	211	11.4	100.0	RSV	756	17.6	99.7
DIP	67	28.6	94.1	SLE	5,675	59.8	99.5
DOB	11	100	100	TPN	1,133	47.4	96.6
DVT	1,507	26.3	33.0	VCR	621	65.7	100.0
EMG	2,277	8.8	38.7	Total	83,491	48.0	92.5

Table 13. The detailed for STC_2 and the quality of STC2 when evaluated on the gold standard set of Set A. STC_2-R and STC_2-P are the recall and the precision respectively.

AW	STC_2-R (%)		STC_2-P (%)	
	CRMap	MetaMap	CRMap	MetaMap
APC	68.8	60.6	84.3	86.3
BSA	9.0	29.9	89.9	89.8
LAM	30.6	16.4	87.5	63.8
MAS	60.7	8.0	98.6	13.6
PVC	25.4	17.5	94.2	56.8
VCR	65.7	100	100.0	99.2

Table 14. The comparing result of two mapping programs: CRMap and MetaMap. Refer to Table 13 for notation of STC_2-R and STC_2-P**.**

The low precision for some abbreviations was caused by the existence of closely related senses. For example, the precision of STC_2 (*EMG*) was 38.7% since three out of four senses are closely related (see Section 6.1.6). *ASP* had two closely related senses: ASP_3 (i.e. *aspartic acid*) and ASP_6 (i.e. *aspartate*). They had relations defined in MRREL, and they also related to 21 concepts in common in MRREL. The precision of STC_2(*ASP*) was 63.8%. All abbreviations that had STC_2 with a low precision had closely related senses.

The quality of STC_2 was related to the quality of conceptual relatives for each sense. For example, a conceptual relative of APC_1 (i.e. *antigen presenting cell*) was *cells*, and the textual variants of *cells* (including *cell*) occurred in many abstracts, therefore our method favored APC_1.

The difference in performance using CRMap and MetaMap indicated the different goals of two programs. The goal of CRMap is to match only conceptual relatives, while the goal of MetaMap is to map every noun phrase in the context to UMLS concepts. MetaMap fails to find conceptual relatives that are preposition noun phrases while CRMap does not have such limitation. For example, MetaMap failed to identify *persistent pulmonary hypertension of the newborn*, which is a sibling of MAS_2 (i.e., *meconium aspiration syndrome*), as relatives of MAS_2 in abstracts that contain it such as "….MAS can easily develop persistent pulmonary hypertension of the new born …". The running time of CRMap is much faster than MetaMap since CRMap only considers conceptual relatives of a specific term.

6.3.3. Experiment III

From the noise tolerance study, we saw that traditional implementation of decision list learning was robust when there was noise in the training set for abbreviations with a skewed sense distribution, and our mixed supervised learning was robust when there are multiple majority senses.

From Experiment I, we observed that for abbreviations with enough instances (over 300 instances) for each sense, almost all WSD classifiers trained on STC_1 achieved a precision of over 97%. For abbreviations with a skewed sense distribution, abbreviations with less than 300 instances for each sense, or abbreviations with two senses that were closely related, implementations of decision list learning achieved the best performance. For abbreviations with a balanced sense distribution for majority senses, our mixed supervised learning achieved the best performance.

From Experiment II, we found that for abbreviations with closely related senses, the derived sense-tagged corpora STC_2 had a low precision, which implied that there was a high level of noise presented in STC_2 for these abbreviations. For other abbreviations, the level of noise was less than 6% (note that different from the noise we introduced to the training set in our noise study, the noise here had a relation with the sense distribution, i.e., rare senses in the gold standard set were usually also rare senses in STC_2).

In this experiment, we assessed the precision of WSD classifiers using the knowledge we had gained in previous experiments but without accessing knowledge presented in individual words.

6.3.3.1. Methods

We split abbreviations into two groups according to the derived sense-tagged corpora and related-ness among senses:

Group 1. contained abbreviations with over 300 instances for each sense in STC_1 or abbreviations with closely related senses including APC, ASP, BSA, DVT, EMG, HSV, IBD, LAM, MAC, RSV, and TPN; where the closely related information among senses was derived from the UMLS (See Appendix E for the detail information about semantic relations extracted from the UMLS);

Group 2. contained other abbreviations.

For each abbreviation in Group 1, we used STC_1 as the training set; for each abbreviation in Group 2, the combination of STC_1 and STC_2 was the training set.

We used our implementation of decision list learning for abbreviations with a skewed sense distribution in Group 1, and our mixed supervised learning for abbreviations with a balanced sense distribution for majority senses, where majority senses were senses that had over 0.5% of occurrences or had over 1% of occurrences with a total occurrence of less than 20. For abbreviations in Group 2, we used traditional decision list learning for abbreviations with a skewed sense distribution and abbreviations with less than 300 instances for each sense, and our mixed supervised learning for abbreviations with a balanced sense distribution for majority senses.

AW	Group	ML	P(%)	AW	Group	ML	P(%)
ACE	2	MSL	96.2	FDP	2	MSL	95.6
ANA	2	TDLL	98.9	HSV	1	MSL	99.7
APC	1	MSL	98.9	IBD	1	TDLL	97.2
ASP	1	MSL	92.9	LAM	1	MSL	82.5
BPD	2	MSL	99.2	LDH	2	TDLL	100
BSA	1	MSL	97.8	MAC	1	MSL	86.7
CAD	2	TDLL	99.1	MAS	2	MSL	99.1
CAT	2	MSL	97.2	MCP	2	MSL	96.3
CML	2	TDLL	94.9	PCA	2	MSL	93.8
CMV	2	TDLL	99.7	PCP	2	MSL	92
CPI	2	TDLL	100	PEG	1	MSL	100
CSF	2	MSL	93.9	PSA	2	MSL	98.3
CVA	2	TDLL	100	PVC	2	MSL	97.7
CVP	2	MSL	100	RSV	1	MSL	96.1
DIP	2	MSL	98.2	SLE	2	TDLL	98.6
DOB	2	TDLL	50	TPN	1	MSL	98.7
DVT	1	TDLL	99.1	VCR	2	TDLL	99.4
EMG*	1	TDLL	21.4	Total	NA		92.0(97.0)

Table 15. The result of Experiment III. The number inside the parentheses in the last row is the overall precision after ignoring EMG.

We used the feature representation which had been used in experiments I and II. We reported the precision of each WSD classifier when evaluated on the gold standard set (note that some instances in $STC_2(W)$ were in the gold standard set).

We reported the performance for each abbreviation as well as the overall performance.

6.3.3.2. Results

Table 15 lists the information about the machine learning algorithm and the precision of the corresponding WSD classifier for each abbreviation. For example, the abbreviation *ACE* belongs to the second group. The supervised learning algorithm used to construct

the WSD classifier for *ACE* was our mixed supervised learning. The precision of the WSD classifier was 96.2% (note that it is lower than the precision (97.5%) reported in experiment I). There were four abbreviations with a precision of less than 92%: *EMG* (21.4%), *DOB* (50%), *LAM* (82.5%), and *MAC* (86.7%). All other abbreviations had a WSD classifier which achieved a precision of over 92%. Three out of four abbreviations with a low precision had closely related senses except *DOB*, which had only two instances in the gold standard set, and therefore the precision does not mean anything. In the training set of *LAM*, LAM_2 was the majority sense while there were no instances in the gold standard set of *LAM* associated with the sense LAM_2. The low precision of *MAC* was also caused by the high degree of ambiguity. There were nine senses present in the training set of *MAC*. Four had over 100 instances for each while all others had less than 40 instances for each.

The overall precision was 92%. After ignoring *EMG*, the overall precision increased to 97.0%.

6.3.4. Evaluation of WSD Classifiers on Instances Extracted from Medical Reports

As we have shown, abbreviations are widely used in the medical reports and usually represent domain-specific terms. Their disambiguation is important for NLP applications in the clinical domain. For example, *PIN* in the MedLEE lexicon has two full forms: *prostatic intraepithelial neoplasia* and *posterior interosseous nerve*, where the former is a kind of disease and the latter is a body location. Both of them carry important clinical information and incorrect interpretations of them are undesirable.

AW	# instances	Gold Standard Sense	Assigned Sense
PCA	4	PCA_5 (posterior cerebral artery)	PCA_4 (posterior communicating artery)
PSA	2	PSA_2 (prostate specific antigen)	PSA_3 (public service announcement)
APC	1	APC_3 (atrial premature complex)	APC_4 (aphidicholin)
CAT	1	CAT_2 (computerized axial tomography)	CAT_4 (combined approach tympanoplasty)
DIP	1	DIP_2 (desquamative interstitial pneumonia)	DIP_4 (distal interphalangeal joint).

Table 16. The detailed information of incorrect sense assignments.

We have shown that a WSD classifier can be constructed by extracting instances of the corresponding full forms from MEDLINE. From a previous study[69], we found that we could not derive enough sense-tagged instances for abbreviations using medical reports only. Can we use the classifiers that are constructed from instances extracted from MEDLINE to disambiguate instances in medical reports? We answered the question through an experiment, where we used the WSD classifiers that were constructed in Section 6.3.3. to disambiguate gold standard instances from the collection of medical reports as described in Section 5.3.

Among 459 instances we extracted from medical reports, 450 (98.0%) were sense-tagged correctly. The detailed information of incorrect sense assignment for 9 instances is listed in Table 16. For example, four instances of *PCA* with the sense PCA_5 were assigned to the sense PCA_4.

From the above experiment, we can see that we can use instances derived from MEDLINE to construct WSD classifiers for the disambiguation task of abbreviations in medical reports. Abbreviations in medical reports seem to have a relatively low level of ambiguity compared to MEDLINE, i.e., they may represent fewer full forms in medical reports than in MEDLINE abstracts. For example, *PSA* as *PSA3* (i.e., *public service announcement*) may not appear in medical reports at all. We can reduce the complexity of WSD classifiers by using those MEDLINE instances, which are associated with senses presenting in the lexical table of a corresponding NLP application, to construct WSD classifiers for the application.

6.3.5. Conclusions

Through a set of experiments presented here, we conclude that for abbreviations with unrelated senses, we can automatically construct WSD classifiers with a high precision (over 97%). If there are enough instances in sense-tagged corpora extracted using synonyms (i.e., STC_1) for each sense, we can construct WSD classifiers using STC_1 only since there is no noise in STC_1. For abbreviations with closely related senses, sense-tagged corpora derived using conceptual relatives (i.e., STC_2) may contain a high level of noise, and therefore we should just use STC_1 to build WSD classifiers for these cases. Traditional decision list machine learning can be used for constructing WSD classifiers when there is noise in the training set; otherwise, our implementation of decision list machine learning is used. Our mixed supervised learning can be used for constructing WSD classifiers for a training set with a balanced sense distribution. In addition, WSD classifiers that are constructed using instances from MEDLINE can be used for the ambiguity resolution in the clinical domain.

6.4. Construction of WSD Classifiers for General Biomedical Terms

Previous experiments have shown that for abbreviations, we can derive sense-tagged corpora that can be used to construct WSD classifiers with good performance. In this section, we first discuss the applicability of our method by providing the number of ambiguous terms in the UMLS that could not use our method to derive sense-tagged instances. We then show that for general English words, such as words in Set B, clustering analysis is necessary and human supervision is required for some cases. There are two experiments to address the issue. The first experiment checked the quality of sense-tagged corpora that were derived for Set B using the UMLS. We then measured the performance of WSD classifiers for words where the gold standard set had a balanced sense distribution with a size of over 15. The second experiment used Set C to show that clustering analysis alone can be used to reduce human annotation costs when constructing a WSD system without using knowledge bases.

6.4.1. Overall Statistics in the UMLS

Among 11,178 concepts that are represented by 4,547 conceptually ambiguous terms in the table AMBIG.SUI, 791 (7.1%) concepts do not have unambiguous synonyms. There are 53,577 unambiguous concept names for the remaining 10,387 concepts with an average of five unambiguous synonyms for each concept. There are 1,262,668 (CUI_1, CUI_2) unique relation pairs defined in MRREL, where CUI_1 is one of the 11,178 concepts that contain an ambiguous concept name. An average of 113 concepts have relations with each CUI_1. There are only 6 (out of 4,547) ambiguous terms, where one concept has no relations defined in MRREL. Combining synonyms and conceptual relatives together, 4 out of 11,178 concepts cannot use our method to derive sense-tagged corpora. The

Term	Concepts	Semantic Types
T1	C0054964	Immunologic Factor/Amino Acid, Peptide, or Protein
	C0657470*	Amino Acid, Peptide, or Protein
Probably	C0332148	Qualitative Modifier
	C0750492*	Idea or Concept
Today	C0310367	Antibiotic, chemical names
	C0750526*	Temporal Concept
Resistance	C0683598	Social Behavior
	C0917925*	Finding

Table 17. Four ambiguous terms in the UMLS that cannot use our two-phase method to derive sense-tagged instances for one sense (flagged with a star).

detailed information about the corresponding ambiguous terms is listed in Table 17. From Table 17, we can see that the two senses of *T1* are closely related; *probably* may not be considered ambiguous in NLP applications since some UMLS semantic classifications are problematic. In addition, *Today* as an antibiotic may be very rare in medical reports compared to its temporal senses according to a physician. Only the term *resistance* may be required to be disambiguated for NLP applications.

6.4.2. Experiment I

We followed the automatic construction of the sense-tagged corpora process, and derived STC_1 and STC_2 for each word in Set B. Detailed information about STC_1 and STC_2 for each sense is listed in Appendix D. Table 18 lists the information of the data as well as the precision and recall of STC_2 when evaluated on the gold standard sets of Set B (note

Word	RC_1+RC_2	STC_1	STC_2	STC_2-R (%)	STC_2-P (%)
ASSOCIATION	1,471	4	1,467	NA	NA
COLD	2,469	1,506	963	28.0	47.3
CULTURE*	2,127	127	309	14.8	100.0
DEGREE*	2,040	40	514	28.3	94.4
DEPRESSION*	2,420	420	1,329	73.3	100.0
DISCHARGE*	2,240	240	880	42.3	91.7
ENERGY	2,221	221	556	8.3	27.6
EXTRACTION	1,998	0	627	6.1	50.0
FAT	2,004	4	905	43.5	81.1
FIT	2,344	344	315	NA	NA
FLUID	2,218	218	1,261	12.3	18.5
FREQUENCY*	2,490	490	607	23.4	88.0
GANGLION*	2,406	409	1,726	72.4	100.0
GLUCOSE	2,342	342	1,140	56.0	90.3
GROWTH	2,246	246	258	7.0	46.7
IMPLANTATION*	2,922	922	765	37.8	90.2
INHIBITION	2,005	5	624	0.0	0.0
JAPANESE	2,234	236	183	8.9	87.5
LEAD	2,009	9	167	50.0	93.3
MAN	2,839	839	1,200	1.1	2.2
MOLE*	2,113	124	139	14.3	100.0
NUTRITION	2,483	483	1,412	30.2	48.5
PATHOLOGY	2,004	4	695	25.5	66.7
PRESSURE*	2,110	110	1,135	24.0	82.1
REDUCTION	2,001	1	333	54.5	100.0
REPAIR	2,242	242	636	26.9	78.3
RESISTANCE	2,017	17	960	0.0	0.0
SCALE	2,003	3	163	0.0	0.0
SECRETION	2,257	257	601	5.7	22.7
SEX	3,090	1,090	930	34.3	87.2
SINGLE*	2,400	400	536	31.0	96.9
STRAINS*	2,093	93	61	8.6	100.0
SURGERY	2,408	408	1,093	40.4	76.9
TRANSIENT*	2,200	200	199	5.2	100.0
TRANSPORT	2,123	123	2,000	30.3	31.0
ULTRASOUND	2,701	701	996	43.0	95.6
WEIGHT	2,406	406	1,643	46.8	78.6
WHITE*	2,524	524	435	27.8	96.2
Total	84,749	11,808	29,763	26.0	65.5

Table 18. The statistics of raw corpora (RC_1 and RC_2) and sense-tagged corpora (STC_1 and STC_2) extracted using our method for words in Set B.

we excluded instances with senses that were not defined in the UMLS, i.e., instances with a sense-tag NONE for the computation of precision and recall). The precision and recall measures for words in Set B are listed in the fourth and fifth columns. From Table 18, we can see that for some words, we cannot derive enough sense-tagged instances using synonyms. Actually, there were 32 senses for each that had less than 10 sense-tagged instances derived using synonyms such as $COLD_5$, $CULTURE_1$ etc. For words with closely related senses such as MAN (refer to Figure 3), the precision of STC_2 was very low (2.2% for man). In the following, we checked the comprehensives of the preliminary sense-tagged corpora for words with unrelated senses while the corresponding STC_2 had a high precision.

We chose the feature representation "e", i.e., collocations, words with the orientation in a window of size 2, and words in a window of size 3 as our features (for a position in the window that contained no words, we used a sign "_"). We removed stop features from the feature representation, where stop features were common features from the top 300 frequent features for all words in Set B with a total of 40 different features including: *and*, *the*, *in*, *or*, *of*, R*/to*, and R*/wa* etc. Each cluster was represented using the top 50 frequent features in that cluster. The similarity of two clusters was computed in the following ways:

- If two clusters share the same collocation, i.e., features $L2L1/w_{L2}\_w_{L1}$, $L1R1/w_{L1}\_w_{R1}$, and $R1R2/w_{R1}\_w_{R2,}$ a score 4/19 was added to the similarity measure;

- If two clusters share the same oriented word, i.e., features L/w_{L2}, L/w_{L1}, R/w_{R1}, and R/w_{R2}, a score 1/19 was added to the similarity measure;

- If two clusters share the same neighboring word (i.e., features $w_{L3,}$ $w_{L2,}$ w_{L1}, w_{R1}, w_{R2}, and $w_{R3),}$ a score 1/38 was added to the similarity measure.

For example, the similarity of Sentence A and Sentence B below is 11/38 since they have the following common features: L1R1/*in_matter* (4/19), R1/*matter* (1/19), and *matter* (1/38).

Sentence A. *An automated brain tissue segmentation procedure was adopted to create anatomical templates to drive feature matching in white matter, gray matter, and cerebral-spinal fluid.*

Sentence B. *We have extended a mathematical model of gliomas based on proliferation and diffusion rates to incorporate the effects of augmented cell motility in white matter as compared to grey matter.*

We used the following threshold vector T: (0.5, 0.45, 0.4, 0.35, 0.3, 0.25). The threshold for the number of clusters in the final clustering was 100. The percentage threshold for the majority sense in each cluster was 90%. The number of clusters and the number of instances in clusters that contained no sense-tagged instances were reported.

Table 19 lists detailed information about the clustering. NC is the number of clusters in the final clustering and NR is the number of instances existing in clusters with no sense-tagged instances. There were two words (i.e., *degree* and *reduction*) that had a final clustering with the number of clusters more than 100. The number of instances in clusters

Word	STC_1+RC_2	NC	NR
CULTURE	2,469	50	18
DEGREE	2,127	119	73
DEPRESSION	2,040	65	18
DISCHARGE	2,420	90	43
FREQUENCY	2,218	95	62
GANGLION	2,490	89	13
IMPLANTATION	2,246	65	9
MOLE	2,839	99	84
PRESSURE	2,004	93	11
REDUCTION	2,110	106	89
SINGLE	3,090	87	54
STRAINS	2,400	65	49
TRANSIENT	2,408	58	44
WHITE	2,406	58	31

Table 19. The result of Experiment I.

that contained no sense-tagged instances was less than 5% of the total for all words, which implied they were comprehensive.

Among 14 words in consideration, after removing the sense None, almost all (over 98%) instances in the gold standard set for eight of them had the majority sense. Two of them (i.e., *mole* and *reduction*) had less than 15 instances in the gold standard set. Two words (i.e. *culture* and *ganglion*) had a precision of over 90% when assigning every instance the majority sense. Only two words (i.e. *implantation* and *white*) had a balanced gold standard set. Using the feature representation "e" with a window size 5, applying Naïve Bayes learning (since the training sets had a balanced distribution), we achieved 85.6% precision for *white* compared to 54.4% when assigning the majority sense to each instance, and 93.9% precision for *implantation* compared to 81% when assigning the majority sense to each instance.

6.4.3. Experiment II

We acquired sentences for each word in Set C from the collection of medical reports. The acquired sentences were divided into a training set and a test set with the ratio 7:3, where the training set was used to build a WSD classifier using clustering algorithm and human supervision, and the test set was used to derive the gold standard set for testing (refer to Section 5.3.3). We used the same feature representation as in Experiment I. The clustering algorithm was applied to each training set. T presented to the clustering algorithm was (0.5, 0.4, 0.3, 0.2, 0.1). For each value of T, we recorded the number of clusters. The clustering process stopped when the number of clusters was less than 200 for a particular value of T, or it stopped after finishing the iteration of T. The human experts were asked to sense-tag an instance randomly selected from each cluster. The sense of that instance was then assigned to each instance in that cluster. We derived a WSD classifier based on the resulting sense-tagged corpus using Naïve Bayes learning with the feature representation "e" and a window size 5. We then ran the WSD classifiers on these 50 instances and computed the precision of the WSD classifiers.

Table 20 lists the result. For the ambiguous word, *cold*, the precision was around 86%. For the others, the precision was 98% or higher.

We feel that the performance of WSD classifiers depends on each individual ambiguous word. Some words are easy to disambiguate, while some words are difficult to disambiguate. There were only two senses of *discharge* and two senses of *lead* were presented in the gold standard set. In the resulting sense-tagged corpus for *discharge*, we found that only 0.5% of the instances had the second sense. Almost 65% of the instances

Word	Training	Testing	No Clusters	P (%)
Cold	1,370	613	106	86
Discharge	16,417	7,040	191	100
Dressing	4,382	1,925	122	98
Lead	4,230	1,815	120	100

Table 20. The result of Experiment II.

had the first sense and 35% had the third sense. In the sense-tagged corpus for *lead*, almost 95% instances had the first sense; 5% had the third sense; less than 0.1% instances with the sense 4 and less than 0.1% instances with the sense 2. We have over 98% accuracy for these three words. We achieved a low accuracy for *cold*. The ratio of different senses of the training set of *cold* is 584:92:307:263:122:2. Five majority senses were present in the gold standard set. Compared to other words, *cold* had a high level of ambiguity and was more difficult to disambiguate.

6.5. Conclusions

In this chapter, we evaluated our two-phase method through several experiments. Based on these experiments, we derived the following conclusions:

The best choice of window sizes depends on certain characteristics of the ambiguous terms. Domain-specific ambiguous terms require a large window such as the whole instance, while general terms require a window of size 2 to 5. Collocations and neighboring words are appropriate features. The best choice of supervised learning

algorithms depends on the sense distribution in the corresponding sense-tagged corpus. For terms with a sense-tagged corpus that is balanced among majority senses, our mixed supervised learning achieves the best performance; for a skewed sense-tagged corpus, traditional decision list learning achieves the best performance when there is noise in the training set; otherwise, our implementation of decision list learning achieves the best performance. Naïve Bayes learning is not an appropriate choice of supervised learning when there are rare senses in the training set.

Our method can be used to construct WSD classifiers for abbreviations automatically with a high precision (around 97%). For other ambiguous biomedical terms, clustering analysis and human supervision are unavoidable if i). the corresponding term has general English senses, ii). there are not enough sense-tagged instances, or iii). senses are closely related. Additionally, WSD classifiers for abbreviations, which are trained on sense-tagged instances derived from MEDLINE, can also be used to disambiguate instances in the clinical domain.

Chapter 7. Applicability Studies

7.1. Requirements of Our Method

There are several requirements needed in order to use our method to derive a WSD classifier for an ambiguous term *W* in an NLP system. One requirement is that the senses of *W* are predefined in the considered NLP system. However, there do exist some rare senses of *W* that may not be captured by the NLP system. For instance, there are two senses of *discharge* in the MedLEE lexicon, which are the discharge procedure and the discharge substance. The sense of *discharge* as electronic discharge appeared in discharge summaries as in the following sentence, "EEG was normal without epileptiform discharge", was not included in the MEDLEE lexicon. Expert-review is unavoidable in order to discover the use of rare senses.

Another requirement is that each sense of *W* can be represented using UMLS concepts. It is not required for extracting STC_1 (i.e., sense-tagged instances extracted using unambiguous synonyms) for abbreviations provided the corresponding full forms are known. Usually, for a biomedical domain NLP system such as MedLEE, which performs clinical information extraction, domain specific ambiguous terms of the system are biomedical terms that most likely can be found in the UMLS.

The method also requires that the corresponding UMLS concepts of each sense of *W* have unambiguous conceptual relatives. We choose the UMLS because it is the most comprehensive biomedical knowledge base and therefore is a valuable resource for WSD.

Our method assumes that all occurrences of *W* in an abstract hold the same sense when using conceptual relatives occurred in the whole abstract for the sense assignment of STC_2 (i.e., sense-tagged instances derived using conceptual related terms occurred in the context). The assumption is obviously true for abbreviations. However, the assumption is required to be verified for general English terms.

Additionally, our method requires that there are enough sense-tagged instances for each sense of *W* and the corpus is comprehensive with respect to senses and genres of contexts. We believe that in order to have an acceptable WSD classifier for W, each sense should have at least hundreds of instances. However for some rare senses of ambiguous words (which may not even be captured by the considered NLP system), it is impossible that enough instances will be captured. For example, the sense of *discharge* as *electronic discharge* occurred only 7 times in the 1998 discharge summary collection (out of 23,651 discharge summaries) while the other two occurred thousands.

We address the applicability of our method through several studies. We first describe the study of the abbreviations in the UMLS. The UMLS coverage of the MedLEE lexicon is presented next. Finally, a feasibility study of the automatic understanding of abbreviations in MEDLINE is discussed.

7.2. The Study of the Abbreviations in the UMLS

7.2.1. Methods

UAExtractor was applied to the 2000 version of the UMLS and extracted a list of 163,666 unique abbreviations pairs.

We first studied the ambiguity presented in the UMLS abbreviation list and hypothesized that ambiguity associated with abbreviations was related to the number of characters in the abbreviations, and that abbreviations with more characters tended to have fewer different full forms than those with fewer characters. We conducted an ambiguity study on a subset of the UMLS abbreviation list: we removed all punctuation marks from each abbreviation, and if the resulting abbreviation had less than 7 characters, it was included in the subset. In our study, if an abbreviation had multiple full forms, it was considered ambiguous. We computed the average number of full forms in the subset. For abbreviations with the same number of characters, we computed the percentage that were ambiguous, the average number of full forms, and the variance.

We then studied the coverage of the UMLS abbreviation list for abbreviations in medical reports. We used the test collection of medical reports (i.e., CMR) to generate two abbreviation sets, I and II. The set I was obtained from CMR by extracting (*AW*, *DOM*, *FF*) tuples, where *AW* is an abbreviation that consists of 2 to 6 characters, *FF* is the associated full form that was defined using parenthetical expressions from reports in domain *DOM*. The set II was obtained by using a program to extract a collection of upper-case words ranging from 2 to 6 characters from mixed-case sentences in the test set. We then obtained a preliminary set of II by selecting 40 words randomly from the collection for each domain, with the restriction that no word appeared in multiple domains (to avoid multiple occurrences of a popular abbreviation in the coverage study). For each word in the preliminary set of B, we randomly selected a mixed-case sentence from a report in the corresponding domain that contained that word. All (*word*, *domain*, *sentence*) tuples were presented to a human expert. For each tuple, the human expert

used all possible sources (the expert's knowledge, abbreviation dictionaries, the WEB, etc) to determine if *word* occurred in the sentence as an abbreviation; if it did, the expert supplied the corresponding *full form*, and the tuple *(word, domain, full form)* became an entry in the abbreviation set II.

For each of the abbreviation sets I and II, we first attempted to automatically map the abbreviation and its full form to the derived abbreviation list. For those that could not be mapped automatically (because of typos in the supplied full forms or different word orders, etc.), we manually searched for them in the UMLS abbreviation list. We computed the ratio of the number of matches against the total number of abbreviations.

We hypothesized that the frequency of abbreviations in the reports was related to the UMLS abbreviation list coverage. We computed ratios associated with five different frequency ranges: I (less than 5), II (between 5 and 10), III (between 10 and 20), IV (between 20 and 50), and V (over or equal to 50). The frequency of an abbreviation is the number of occurrences of that abbreviation in the test set. The ratio for each range consisted of abbreviations that had the correct full forms in the UMLS divided by the total number in that range.

7.2.2. Results

In the ambiguity study, there were 16,855 abbreviations in the set. We found that 33.1% of them had multiple full forms. The average number of full forms for abbreviations with less than 7 characters was 2.28. Table 21 lists the results with respect to the number of characters: *Len* is the number of characters in the abbreviation; *Num* is the number of

Len	Num (%)	Avg	V
1	26 (100)	52.6	6.48
2	596(81)	10.9	0.59
3	4,137(54)	3.05	0.06
4	5,051(27)	1.64	0.03
5	3,777(21)	1.41	0.01
6	3,268(20)	1.33	0.02

Table 21. The ambiguity study results with respect to the number of letters in the abbreviations

Domain	Num	I (%)	II (%)
Neurology	2,758	13(46)	40(70)
Pathology	102,933	132(64)	33(70)
Discharge	23,651	86(72)	33(55)
Ob/Gyn	12,198	3(0)	29(59)
Radiology	306,587	41(78)	33(82)
GI Endoscopy	6,121	3(67)	40(75)
Cardiology	123,799	21(67)	37(76)
Surgery	39,333	65(62)	25(56)

Table 22. The UMLS abbreviation coverage results with respect to the domain

abbreviations; the number in parentheses is the percentage of ambiguous abbreviations; *Avg* is the average number of full forms; *V* is the variance of the number of full forms. There were 364 tuples in the set I; 241 (66.2%) were mapped to the UMLS abbreviation list. The abbreviation set II contained 270 tuples (84.4% of the preliminary set of II); 185 (68.5%) were mapped to the UMLS abbreviation list. Table 22 lists the results for the

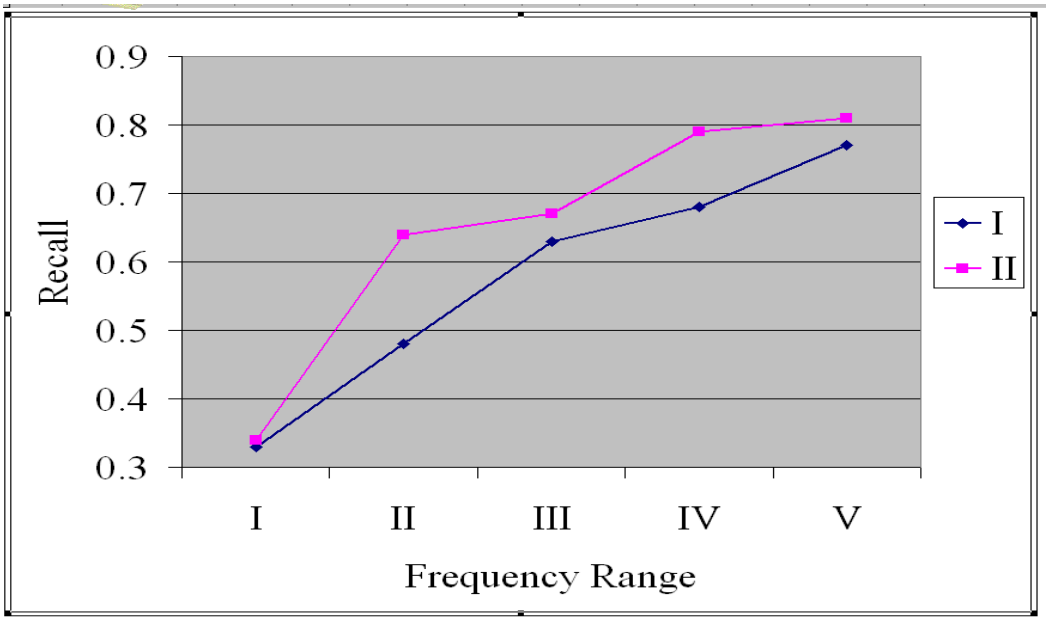


Figure 10. The UMLS abbreviation coverage result with respect to frequency.

sets I and II with respect to the domain: *Num* represents the number of reports; the number in parentheses is the percentage of abbreviations that have matches in the UMLS abbreviation list. Figure 10 lists the results for the sets I and II with respect to the five frequency ranges. The X-axis represents the range and the Y-axis represents the ratio. Only 30% of the abbreviations occurring less than 5 times in the medical reports were found in the UMLS abbreviation list whereas 80% of the abbreviations occurring more than 50 times were found.

7.3. The UMLS Coverage of the MedLEE Lexicon

Since senses of lexical entries in the MedLEE lexicon are represented using (s, t) pairs, where s is the semantic type and t is the target form, the mapping of the MedLEE lexicon to the UMLS includes not only term mappings, but also semantic type mappings. We

studied the UMLS coverage of the MedLEE lexicon by generating a lexicon for MedLEE using the UMLS.

7.3.1. Methods

We first determined the semantic categories in the UMLS semantic network that were appropriate to use for the purpose of generating entries for the MedLEE lexicon without manual intervention. A separate file was created for each category containing all concept names that corresponded to that category. A program was written to check whether the concept names in each of those files had a matching lexical entry in MedLEE's lexicon. The semantic categories (as specified by MedLEE) of all concept names that matched were recorded and counted. The majority MedLEE semantic category in each file was considered as the appropriate category for the corresponding UMLS category. Those files that did not have the majority MedLEE semantic category were excluded from further consideration.

The concept names in files that remained under consideration were then to used to create a UMLS-based MedLEE lexicon. There were two sets of concept names that were kept for further consideration: concept names that were entries in the Specialist Lexicon after a normalization process and concept names that appeared in the collection of medical records (i.e., CMR). The normalization process removed the symbols NOS, certain punctuation marks, and other symbols (i.e. <1>) were removed, and all letters were mapped to lower case. A following heuristic was applied to check the existence of a concept name in CMR:

- All one-grams and bi-grams with occurrences of 5 or more were extracted from CMR.

- For each concept name in the META, we first extracted its one-grams and bi-grams. If all of those grams had corresponding entries in CMR, then such name was considered presenting in CMR.

- The resulting UMLS semantic lexicon was compared with a subset of the MedLEE clinical lexical entries (with 10,998 entries)[21].

7.3.2. Results

There were 252,134 different lexical candidates extracted from the META. 112 out of 134 UMLS semantic categories were automatically matched to the MedLEE semantic categories. 6,017 (54.7%) MedLEE lexical entries were mapped to the same entries in the UMLS semantic lexicon; 1,379 (12.5%) had the same lexical entries but different semantic categories; 3,704 (33.7%) had no matches.

We studied the UMLS coverage of the MedLEE lexicon through building a MedLEE lexicon using the UMLS. Using a string matching method such as MetaMap may result a better coverage but may not reflect the true UMLS conceptual coverage of the MedLEE lexicon. The measures provided here were very conservative. Part of the resulting lexicon has been evaluated using an automatic system that computed the risk classification of patients with community-acquired pneumonia[34] and achieved good result but were significantly different from the MedLEE lexicon.

[21] The 2000 versions of the UMLS and the MedLEE lexicon were used.

We checked those MedLEE lexemes that could not be mapped to the UMLS semantic lexicon and found there were several causes:

- Different granularities: for example, there is no mapping for *arteriogram* in the 2000 version of the UMLS, but there were entries such as *renal arteriogram*, or *kidney arteriograms* etc[22];

- Textual variants: for example, *aortic cannula* in the MedLEE has a mapping in the UMLS (i.e., C0179557, *cannulae, aortic*).

Extracting the occurrence information from a larger corpus and considering the frequency of fixed sub-structures in the META may result a lexicon that is comparable to the MedLEE lexicon.

7.4. Automatic Understanding of Abbreviations in MEDLINE

As we have already known, abbreviations are everywhere in the biomedical domain [63]. The understanding of abbreviations in a document is often a difficult task for computer systems. The abbreviation problem has been shown to affect knowledge-based systems, such as information retrieval systems and information extraction systems in biomedicine[7;31;82].

First, a method to associate an abbreviation to its corresponding full form (also termed as expansion or definition) in the context is needed, with an assumption that the authors define the less well-known abbreviations when they are first introduced in a specific domain. Secondly, well-known abbreviations are not always defined in documents. In order to understand these, an abbreviation database that lists abbreviations together with

their senses needs to be built and updated periodically. However, manually constructing a database is time-consuming. In addition, manual maintenance and further extension are increasingly complex. But constructing an abbreviation database automatically by matching abbreviations with their full forms in the document requires a method to group textual variants together and a method to link them to the proper sense. Finally, abbreviations are highly ambiguous. The number of characters that form an abbreviation is limited, and abbreviations are usually short. With the rapid growth of the use of abbreviations, one abbreviation may represent dozens of senses. A method to resolve the sense ambiguity is needed.

We have shown that the UMLS contained many abbreviations together with their full forms, and the ambiguity of abbreviations could be resolved using an automated method if the corresponding full forms occurred frequently.

In this section we address the following issues with respect to abbreviations in MEDLINE abstracts by using three-letter abbreviations: can we build an abbreviation knowledge base from MEDLINE abstracts? If yes, what is the UMLS concept coverage, what is the average number of textual variants for each sense, how ambiguous are the abbreviations, and what is the role of the frequency of the senses?

7.4.1. Background and Related Work

There are several studies on matching abbreviations to their corresponding full forms in documents. Taghva et al.[107] developed an algorithm that considers strings of from 3 to 10 uppercase letters as acronyms, and looks for candidate full forms in windows twice

---

[22] The 2001 version of the UMLS had an entry for arteriogram

the length of number of the acronym located immediately before or after the acronym. Larkey et al.[63] implemented a Web server for abbreviations, where abbreviations and their full forms were gathered automatically from a large number of Web pages. Yoshida and colleagues[125] built a workbench for the construction of a protein-abbreviation dictionary. Yu and colleagues[126] developed a program to extract full forms of abbreviations from full articles. All the above studies achieved precision of over 97% when matching abbreviations to their full forms in documents. However, none of them provide a detailed analysis of characteristics of abbreviations with respect to senses.

In order to pursue our study, we developed a method, PW3, based on Larkey's method, for three-letter abbreviations where the associated full forms were defined in parenthetical expressions. We did not conduct an evaluation of PW3 since our primary goal was to address the UMLS coverage of three-letter abbreviations in MEDLINE.

There are several reasons we used three-letter abbreviations for the coverage study. First, a method for pairing three-letter abbreviations with their full forms is easy to develop and has high precision according to Larkey et al.[63]. Secondly, a preliminary investigation showed that three-letter abbreviations were the most frequent in MEDLINE abstracts. In addition, unlike two-letter abbreviations, which can have several dozens of full forms, the ambiguity of three-letter abbreviations is moderate, whereas most abbreviations with more than 3 letters are not ambiguous.

7.4.2. Methods

The study contained several steps. The first step derived a collection of (*AW*, *FF*, *FREQ*) tuples, where *AW* is a three-letter capitalized text string, *FF* is its associated full form

derived from abstracts using PW3 (the program we developed to pair three-letter abbreviations with full forms), and *FREQ* is the number of abstracts from which PW3 derived the pair. The second step mapped the full forms to the UMLS using FFMap, which is based on MetaMap. The third step grouped similar full forms for the same abbreviation together using FFGrouper (a program to group similar full forms together according to several normalization criteria). The fourth step assessed results, where full forms in the same group were treated as textual variants of the same sense. In the following, we describe PW3, FFMap, and FFGrouper in detail. The assessment method is presented last.

PW3: a matching method for three-letter abbreviations AW

PW3 is designed to search for a possible full form from candidate text strings within a window size 6 at the left side of a parenthetical expression "(*AW*)". It applies several full form patterns of *AW* and three groups of words that can be ignored when matching patterns.

The full form patterns include the following several cases:

- Three letters of *AW* are initial letters of three different words in the right order: e.g. *minimum alveolar anesthetic concentration (MAC),*

- Two letters of *AW* are initial letters of two words and the remaining one appears in one of these two words in the right order followed by at least three letters*:* e,g. *procoagulant activity(PCA)* or *indirect immunofluorescence (IIF),*

- Three letters of *AW* appear in one word where the first one is the initial letter of the word and remaining two appear in the right order: e.g. *carboxymethyllysine (CML).*

PW3 has an additional pattern for potential chemical abbreviations, where *AW* is considered to be a chemical abbreviation if a candidate string contains a number (or a comma or right parenthesis) followed by a non-space letter or a left parenthesis preceded by a non-space letter:

- Two letters of *AW* are initial letters (or following punctuations and numbers) and the remaining letter appears in the corresponding candidate string: e.g. *n-6-(delta-2-isopentenyl)adenine (IPA)*.

The three groups of words which can be ignored when matching patterns are pre-inclusion words (i.e. a word at the beginning of a full form, such as *department, office* etc), post-inclusion words (i.e. a word at the end of a full form, such as *acid, protein, enzyme* etc), and stop words (a word in the middle of a full form, such as *of, for, and, the* etc). PW3 allows one pre-inclusion word, one post-inclusion word, one other word, or two stop words in a full form. The number of words in the full form is at most 6. These three groups words were learned from the three-letter abbreviations in the SPECIALIST manually.

FFMap

FFMap is based on MetaMap. FFMap uses the following subset of concept names in the Metathesaurus: chemical names, concept names that contain less than 7 words after a normalization process, and full forms obtained by executing the UMLS extraction

program. All concept names are normalized by removing some patterns (e.g. *As* – in *As – Arsenic* and *(WS)* in *West syndrome (WS)*), changing to lower-case, and replacing certain punctuation by blanks. In addition, FFMap applies a synonym-like set, which contains pairs (w_1, w_2), where w_1 and w_2 are different words in two concept names of the same UMLS concept. For example, (*hepatic*, *liver*) is a synonym-like pair, which is derived from two concept names of C0009714, *congenital hepatic fibrosis* and *congenital fibrosis liver*.

The input to FFMap is a pair (*AW*, *FF*) and the output is (*AW*, *FF*, *CUI*, *PN*, *MODE*), where *AW* is an abbreviation, *FF* is a full form, *CUI* is the resulting concept identifier, *PN* is the preferred name of that concept, and *MODE* is the matching mode, which can have one of the following four values:

Exact -- a concept name of *CUI* is identical to *FF*, e.g., *(BAL, bioartificial liver, C0336562, artificial liver, exact)*;

SPECIALIST-normalized--a concept name of *CUI* is identical to *FF* when normalized using the SPECIALIST and word order is disregarded, e.g., *(CLD, chronic liver diseases, C0341439, chronic liver disease, SPECIALIST-normalized)*;

Stemmed--a concept name of *CUI* is identical to *FF* when stemmed and disregarding word order, e.g., *(BHC, benzenehexacarboxylic, C0105581, benzenehexacarboxylate, stemmed)*;

Synonym-like-replacement--a concept name of *CUI* is identical to *FF* after replacing one word in *FF* using a synonym-like set and ignoring word order, e.g., *(HFT, hepatic function test, C0023901, liver function tests, synonym-like-replaced)*.

FFGrouper: grouping full forms

FFGrouper is a program to group similar full forms of the same abbreviation together. For an abbreviation *AW*, each of its full forms consists of a group. FFGrouper groups similar groups of *AW*, subsequently using the following normalization phases:

Group by ignoring punctuation: after removing punctuation, if a full form in one group is the same as a full form in another group, the two groups are merged. For example, three different full forms for *IGS*: *immunogold staining*, *immuno gold staining*, and *immuno-gold staining* are merged into the same group.

Group using the SPECIALIST: after normalizing using the SPECIALIST, if a full form in one group is identical to a full form in a different group, the two groups are merged. For example, the group for *IGS* containing *immuno-gold stain* is merged with the group containing *immuno-gold staining*.

Group by ignoring stop words, word order, correcting typos and expanding abbreviation: two groups are merged together if after ignoring word order, two full forms (one from each):

- are identical after ignoring stop words (e.g., the group for *IMT* containing *intima-media thickness* is merged with the group containing *intima and media thickness*);
- differ in one type-error operation, i.e., replacement, transposition, insertion and deletion (e.g., the group for *IGR* containing *insect grwoth regulator* is merged with the group containing *insect growth regulator*);

- differ in a two-letter abbreviation and its full form (e.g. the group for *MIF* containing *micro-if* is merged to the group containing *micro-immunofluorescence*).

Assessment

PW3 was executed for all MEDLINE abstracts up to December 2001. For each abbreviation *AW*, the number of abstracts that contained the parenthetical expression "(*AW*)" as well as the number of abstracts that contained *AW* with full forms found by PW3 was measured.

We evaluated FFMap using MetaMap. We used MetaMap to get mappings for all full forms with the following options: a) Adj/Noun Derivational Variants, b) No Acronym/Abbreviation variants, c) Stop larger N, d) Ignore word order, e) Truncate candidate strings, and f) Strict mode. If the resulting mappings were single concepts with a relatively high matching score, the mappings were considered as appropriate mappings. For example, if a full form of IGR *intergenic region* was mapped to a single concept C0887859 with a score 1000, the mapping result was *(IGR, intergenic region, C0887859, 1000)*. The intra-agreement of the two systems was computed. In addition, we manually checked mapping results for full forms that occurred more than 200 times, and for which MetaMap either did not have the mappings or had mappings that were different from FFMap.

After grouping full forms using FFGrouper, we further grouped full forms according to the mapping results since two groups with the same concept identifier have the same sense. For example, the group of *IHD* that contains *ischemic cardiac disease* is merged to

the group that contains *ischemic heart disease* since these two full forms are concept names of the same concept.

We computed the average number of variants for each group, and the number of groups with full forms having mappings associated with eight frequency thresholds: 1, 5, 10, 50, 100, 200, 500, and 1000, where the frequency of each group is the summation of occurrences of all full forms in that group. The ambiguity was measured considering the number of groups of each abbreviation with respect to five frequency thresholds: 1, 2, 5, 10, and 100.

7.4.3. Results

We excluded four capitalized text strings (i.e., III, VII, XII, XXI) from the result since they usually represented numbers. Among 4,839,200 unique occurrences of the parenthetical expression "(*AW*)", PW3 extracted 1,793,479 (*AW*, *FF*) pairs, where 206,964 unique (*AW*, *FF*, *FREQ*) tuples were derived (*FREQ* is the number of abstracts that have *FF* as full form for *AW*). The tuples with a *FREQ* value larger than 10,000 were:

- *(PCR, polymerase chain reaction, 19,067),*
- *(HIV, human immunodeficiency virus, 15,232),*
- *(MRI, magnetic resonance imaging, 12,855)*
- *(LPS, lipopolysaccharide, 12,816)*
- *(PKC, protein kinase c, 11,162)*
- *(CNS, central nervous system, 10,497)*
- *(CSF, cerebrospinal fluid, 10,710).*

For the 35,981 full forms with mappings found by both FFMap and MetaMap, the intra-agreement between the two systems was 99.6%. MetaMap matched 1,280 full forms for which FFMap failed to find a match. FFMap matched 14,230 full forms for which MetaMap failed to find a match. Among 39 full forms that we manually checked, 36 full forms were correct mappings (including 28 exact mappings for chemical names).

Among 50,211 full forms with mappings found by FFMap, 31,223 of them were exact mappings, 13,871 were SPECIALIST-normalized mappings, 880 were stemmed mappings and the remaining 4,237 mappings were associated with synonym-like-replacement.

The number of groups was 155,302. The average number of variants for each group was 1.33. The group with the largest number of variants was *(FDG, 18f-fluorodeoxyglucose)* with 170 variants. 23.5% of the groups had full forms with mappings found by FFMap, and covered 77.8% of the occurrences. Table 23 lists the results with respect to different thresholds: FTV is the frequency threshold value, NG is the number of groups, FREQ is the number of overall occurrences, AV is the average number of variants, PG is the percentage of groups with mappings, and PO is the percentage of occurrences with mappings. For example, after disregarding groups with occurrences of less than 500, there were 505 groups, with an overall frequency of 872,035; the average number of variants for each group was 13.66; 96.2% of the groups had full forms with mappings found by FFMap, which covered 98.0 of the total occurrences.

FTV	NG	FREQ	AV	Mapping	
				PG (%)	PO(%)
1000	247	695,902	17.23	98.8	99.1
500	505	872,035	13.66	96.2	98.0
200	1,168	1,078,704	10.06	88.9	95.4
100	2,187	1,221,825	7.74	82.8	93.1
50	3,850	1,336,424	6.01	77.1	91.2
10	13,445	1,534,198	3.45	59.3	86.5
5	23,841	1,601,503	2.72	50.8	84.6
1	155,302	1,793,479	1.33	23.5	77.8

Table 23. The number of variants and the UMLS coverage with respect to eight thresholds.

FTV	NA	NG	POA (%)	PANA (%)	AAMB
1	11,328	155,302	99.6	81.2	16.6
2	9,299	64,338	98.2	74.5	8.95
5	6,767	23,841	94.0	64.6	4.91
10	5,297	13,445	87.7	55.4	3.78
100	1,683	2,187	42.3	22.0	2.36

Table 24. The ambiguity assessment result with respect to five thresholds.

Among 11,328 different abbreviations, 81.2% were ambiguous and covered 99.6 of the total occurrences, with an average of 16.6 groups. Table 24 lists the results with respect to five thresholds: FTV is the frequency threshold value, NA is the number of abbreviations, NG is the number of groups, POA is the percentage of occurrences of ambiguous abbreviations, PANA is the ratio of the number of abbreviations that are

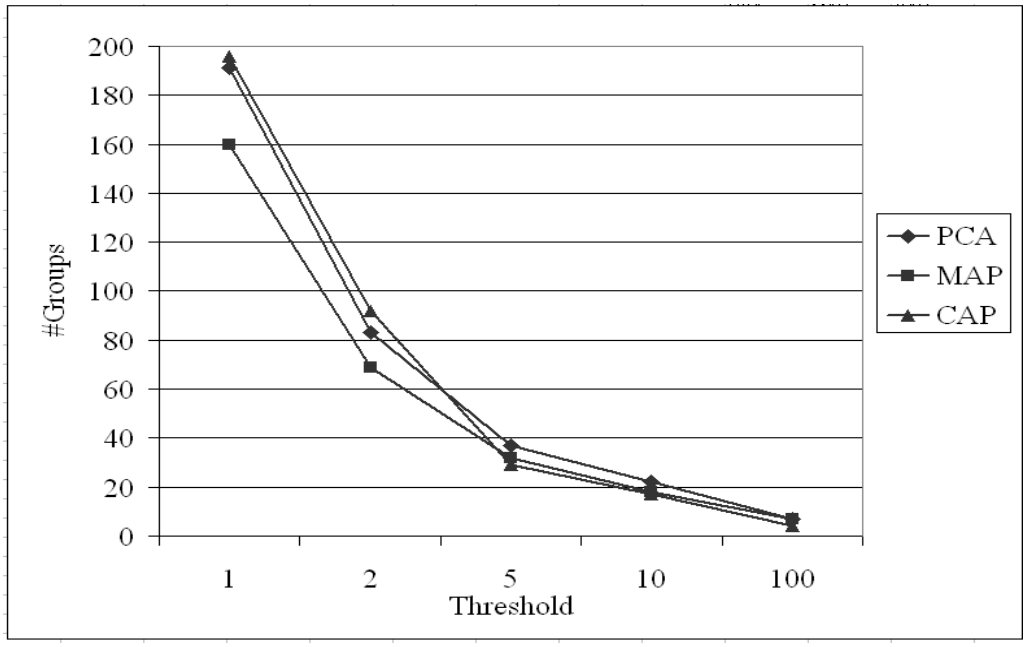


Figure 11. The ambiguity in relation with five frequency threshold values for the most ambiguous three abbreviations.

ambiguous to the total, and AAMB is the average ambiguity for ambiguous abbreviations, i.e., abbreviations with more than one group. For example, after disregarding groups with occurrences of less than 5, there were 6,767 abbreviations with 23,841 groups; 64.6 of the abbreviations appeared in more than one group, with an average of 4.91 groups for ambiguous abbreviations. The three most ambiguous abbreviations were CAP, MAP, and PCA. Figure 11 shows the ambiguity in relation with five frequency thresholds for these three ambiguous abbreviations.

7.4.4. Discussion

FFMap, which is based on MetaMap, is comparable to it: the two systems only disagreed on 0.4 of the mappings. We originally planned on using MetaMap exclusively to find mappings. However, MetaMap with the above-specified options did not give mappings

for most chemical names (especially those with numbers or punctuations). We believe MetaMap can give more mappings using different options.

In this study, we grouped similar full forms of the same abbreviation together to assess coverage and ambiguity. We believed similar full forms would have similar senses. An abbreviation knowledge base can be built automatically from MEDLINE abstracts by utilizing automatic methods to pair abbreviations with full forms. We found that frequency of senses plays an important role in the assessment:

- the UMLS coverage: those with higher frequency were more likely to have a mapping concept. For example, from Table 23, we can see that 23.5% of the senses with occurrences of at least 1 were mapped to the UMLS; while for senses with occurrences of at least 100, 82.8% had mappings.

- the number of textual variants: senses with higher frequency had more textual variants, which follows Zipf's law, i.e., senses with high frequency tend to have many synonyms. For example, from Table 23, we can see that senses with occurrences of at least 1 had an average of 1.33 variants; while senses with occurrences of at least 100 had an average of 7.74 variants.

- the ambiguity of abbreviations: abbreviations were less ambiguous when ignoring rarely occurring senses. For example, from Table 24, we can see that 81.2% of abbreviations were ambiguous with an average of 16.6 senses; after ignoring senses with a frequency of less than 10, 55.4% of abbreviations were ambiguous with an average of 3.78 senses.

We did not evaluate FFGrouper because of a lack of a gold standard. Some full forms with different senses were incorrectly merged together. For example, (*IMN, intramedullary nail*) and (*IMN, intramedullary nailing*) were merged together by FFGrouper, but had different concept identifiers: C0348001 for the former one and C0021885 for the latter. About 630 out of 155,302 groups were mapped to different concept identifiers by different full forms in the same group.

7.5. Conclusions

We studied the applicability of our method through a set of studies that address the coverage of the UMLS and the ambiguity of abbreviations.

Results demonstrated that abbreviations were very ambiguous. The ambiguity of an abbreviation depended on the number of letters it contained: ones with fewer characters were more ambiguous. The ambiguity of an abbreviation also depended on the frequency it occurred in a free-text collection: ones with higher frequency were more ambiguous.

The UMLS covered around 80% of frequent abbreviations either from medical reports or MEDLINE abstracts. There were 54.7% of MedLEE lexicon entries that were automatically mapped to the UMLS with the correct associated semantic categories.

We studied the feasibility of automatic understanding of abbreviations by studying several characteristics of three-letter abbreviations in MEDLINE abstracts. We found that automatic understanding of abbreviations is feasible for frequently occurring abbreviations.

Chapter 8. Future Work and Conclusions

8.1. Future Work

Future work of the current research includes the following several directions.

First, we plan to integrate the WSD system with the MedLEE system. Ambiguous abbreviations have shown to impact the performance of the MedLEE system. As we have shown, WSD classifiers for abbreviations can be automatically constructed using MEDLINE. The MedLEE lexical table contains a subset of abbreviations that experts have already derived a set of full forms for them from medical reports. We will use those given full forms to derive sense-tagged instances from medical reports and MEDLINE abstracts. If there are enough instances, we can use our implementation of decision list learning as the supervised learning algorithm, where the resulting list can be transformed to PROLOG predicates.

In the present work, we haven't used related concepts in the context for the sense assignment of ambiguous terms in medical reports. We plan to derive sense-tagged corpora from medical reports using concepts in the context and study the relation between window sizes and the performance.

If feasible, we plan to study the relation between different version of the UMLS and the ambiguity resolution of terms. As we know that the UMLS keeps adding new concepts and merging similar concepts. In addition, new conceptual relations keep adding to the conceptual network. We have said that the UMLS becomes more and more appropriate

for the ambiguity resolution. However, we did not have a thorough study about it because the availability of resources (i.e., different versions of the UMLS).

8.2. Conclusions

A WSD system that resolves sense ambiguities is essential for the improvement of NLP applications in the biomedical domain. Several preliminary WSD methods for NLP applications in the domain were based on handcrafted rules, which were often incomplete and un-scalable. Supervised machine-learning techniques have been used to construct WSD classifiers automatically from sense-tagged corpora. However, manual sense-annotation of a corpus is also a manual task.

In this dissertation, we proposed a two-phase WSD method where the first phase derives a sense-tagged corpus automatically (may be followed by a human supervision process using clustering analysis techniques) and the second phase builds a WSD classifier through supervised learning on the derived sense-tagged corpus automatically.

Our two-phase WSD method can be used to derive WSD classifiers for abbreviations with a set of known full forms with a high precision (around 97%) without the requirement of clustering analysis. WSD classifiers trained on sense-tagged instances, which are automatically derived from MEDLINE, can be used to disambiguate instances in the clinical domain.

Our two-phase WSD method can also be applied to derive WSD classifiers for a majority of ambiguous UMLS biomedical terms. The derived WSD classifiers achieve a high precision for terms with a set of unrelated senses. However, human supervision is

unavoidable for terms with close-related senses and for general English terms; and clustering analysis can reduce human annotation cost dramatically.

As a unique comprehensive and large size machine-readable dictionary, the UMLS provides a good coverage of biomedical concepts, which is suitable to be used as knowledge bases in this dissertation.

In summary, the contributions of the dissertation include the following:

- This is the first systematic WSD work in the biomedical domain. Researchers in the computational linguistics field debate the soundness of treating WSD as a classification task as part of speech tagging, and the feasibility of building a universal WSD system[55;116].

- Large-scale evaluations of WSD systems are typically impeded by the lack of a gold standard set[56;57]. We provide a method for automatic evaluation of our WSD system using abbreviations.

- We provided a thorough comparison study of different supervised WSD classifiers with four variables: type of ambiguous terms, feature representation, supervised learning algorithm, and window size. We also compared the noise tolerance of different supervised learning algorithms.

- Our implementation of decision list learning, which separates features that occur with only one sense from other features, has a better performance than traditional implementations of decision list learning, which do not distinguish these two, when there is no noise in the training set.

- Traditional WSD implementations of Naïve Bayes learning do not distinguish rare senses from majority senses in the training set. We divided these two and proposed a mixed supervised learning algorithm that combines a Naïve Bayes classifier with an instance-based classifier using a local similarity measure (i.e., the computation of the similarity between two instances is only based on features of these two instances).

- This is the first large-scale WSD work that combines sense-tagged corpora derived using machine-readable dictionaries with supervised machine learning techniques. Previous WSD work isolates these two[4;77;105].

- We discovered that the best choice of window size depends on certain characteristics of the ambiguous terms. Domain-specific ambiguous terms require a large window such as the whole instance, while general terms require a window of size 2 to 5. The best choice for a supervised learning algorithm depends on the sense distribution in the corresponding sense-tagged corpus. For terms with a sense-tagged corpus that is balanced among majority senses, our mixed supervised learning achieves the best performance; for a skewed sense-tagged corpus, traditional decision list learning achieves the best performance when there is noise in the training set; otherwise, our implementation of decision list learning achieves the best performance.

- We developed a clustering algorithm that can handle a large number of instances with a large number of features without the requirement of a pre-determined fixed number of clusters. Most existing clustering algorithms are optimized and suffer from either a speed or space problem [108]. We sacrifice a little bit of the clustering quality to solve these problems.

Appendix

Appendix A. Detailed sense definitions for Set A

Sense	CUI	Full Form
ACE1	C0001044	acetylcholinesterase
ACE2	C0022709	angiotensin converting enzyme
ACE3	C0050385	doxorubicin cyclophosphamide
ACE4	C0108844	doxorubicin cyclophosphamide etoposide
ACE5	C0286421	amsacrine cytarabine etoposide
ACE6	C0304721	adrenocortical extract
ACE7	C0473028	antegrade colonic enema
ANA1	C0002463	american nurses association
ANA2	C0003243	antinuclear antibody
ANA3	C0027385	alpha naphthylesterase
APC1	C0003315	antigen-presenting cells
APC2	C0032580	adenomatous polyposis coli
APC3	C0033036	atrial premature complexes
APC4	C0085171	aphidicholin
APC5	C0809732	activated protein c
ASP1	C0003431	antisocial personality
ASP2	C0003993	asparaginase
ASP3	C0004015	aspartic acid
ASP4	C0038013	ankylosing spondylitis
ASP5	C0052546	aspartylglycine
ASP6	C0085845	aspartate
BPD1	C0006012	borderline personality disorder
BPD2	C0006287	bronchopulmonary dysplasia
BPD3	C0552399	biparietal diameter
BPD4	C0729200	blood pressure decrease
BSA1	C0005902	body surface area
BSA2	C0036774	bovine serum albumin
CAD1	C0010068	coronary artery disease
CAD2	C0170509	cyclophosphamide dacarbazine doxorubicin protocol
CAD3	C0280573	cytarabine daunorubicin
CAD4	C0282308	chronic actinic dermatitis
CAD5	C0669173	caspase activated dnase
CAT1	C0008169	chloramphenicol acetyltransferase
CAT2	C0040405	computerised axial tomography
CAT3	C0041207	common arterial trunk
CAT4	C0395734	combined approach tympanoplasty
CAT5	C0908142	cool associated tyrosine

Sense	CUI	Full Form
CML1	C0023473	chronic myeloid leukemia
CML2	C0301896	cell mediated lympholysis
CMV1	C0010825	cytomegaloviruses
CMV2	C0190084	closed mitral valvotomy
CMV3	C0285131	cisplatin methotrexate vinblastine
CMV4	C0419012	controlled mandatory ventilation
CPI1	C0009825	consumer price index
CPI2	C0451055	california personality inventory
CPI3	C0671646	cyclopropapyrroloindole
CSF1	C0007806	cerebrospinal fluid
CSF2	C0009392	colony stimulating factors
CSF3	C0072454	cytostatic factor
CSF4	C0893357	competence and sporulation factor
CVA1	C0038454	cerebral vascular accident
CVA2	C0054889	cyclophosphamide vincristine doxorubicin
CVP1	C0056633	cyclophosphamide vincristine prednisone
CVP2	C0280556	cisplatin cyclophosphamide etoposide
CVP3	C0520454	central venous pressure
DIP1	C0057737	diazenedicarboxylic acid bis n methylpiperazide
DIP2	C0238378	desquamative interstitial pneumonia
DIP3	C0833631	distal interphalangeal
DOB1	C0231796	disorder of breathing
DOB2	C0301362	bromdimethoxyamphetamine
DOB3	C0421451	date of birth
DVT1	C0149871	deep vein thrombosis
DVT2	C0151950	deep vein thrombophlebitis
EMG1	C0004903	exomphalos macroglossia gigantism
EMG2	C0013839	electromyography
EMG3	C0180677	electromyographs
EMG4	C0392125	electromyogram
FDP1	C0016763	fructose diphosphate
FDP2	C0060663	formycin diphosphate
FDP3	C0163275	fibrinogen degradation product
FDP4	C0224261	flexor digitorum profundus
FDP5	C0851147	followup drinker profile
HSV1	C0206558	herpes simplex virus
HSV2	C0242529	highly selective vagotomy
IBD1	C0021390	inflammatory bowel diseases
IBD2	C0022104	irritable bowel syndrome
LAM1	C0065041	lipoarabinomannan
LAM2	C0205274	laminated
LAM3	C0206621	lymphangiomyomatosis
LAM4	C0282400	leukocyte adhesion molecule
LAM5	C0751674	lymphangioleiomyomatosis

Sense	CUI	Full Form
LDH1	C0022917	lactate dehydrogenase
LDH2	C0851148	lifetime drinking history
MAC1	C0009545	membrane attack complex
MAC2	C0024432	macrophage
MAC3	C0026914	mycobacterium avium complex
MAC4	C0083360	methotrexate dactinomycin cyclophosphamide
MAC5	C0332573	macula
MAC6	C0451273	macandrew alcoholism scale
MAC7	C0453947	mackintosh
MAC8	C0497677	monitored anesthesia care
MAC9	C0582645	mental adjustment to cancer
MAS1	C0016065	mccune albright syndrome
MAS2	C0025048	meconium aspiration syndrome
MAS3	C0451273	macandrew alcoholism scale
MCP1	C0024994	2 methyl 4 chlorophenoxyacetic acid
MCP2	C0025525	metacarpophalangeal joint
MCP3	C0025843	multicatalytic protease
MCP4	C0025853	metoclopramide
MCP5	C0282566	monocyte chemoattractant protein
MCP6	C0285488	membrane cofactor protein
PCA1	C0030131	para chloroamphetamine
PCA2	C0030625	passive cutaneous anaphylaxis
PCA3	C0078944	patient controlled analgesia
PCA4	C0149559	posterior communicating artery
PCA5	C0149576	posterior cerebral artery
PCA6	C0261200	pedal cycle accident
PCA7	C0411287	percutaneous angioplasty
PCA8	C0429865	principal component analysis
PCA9	C0474316	appt canceled by patient
PCP1	C0009414	posterior colpoperineorrhaphy
PCP2	C0030135	p chlorophenylalanine
PCP3	C0030855	pentachlorophenol
PCP4	C0031381	phencyclidine
PCP5	C0032305	pneumocystis carinii pneumonia
PEG1	C0032483	polyethylene glycols
PEG2	C0176751	percutaneous endoscopic gastrostomy
PSA1	C0003872	psoriatic arthritis
PSA2	C0138741	prostate specific antigen
PSA3	C0687688	public service announcement

Sense	CUI	Full Form
PVC1	C0032624	polyvinylchloride
PVC2	C0151636	premature premature complex
PVC3	C0280556	cisplatin cyclophosphamide etoposide
RSV1	C0035236	respiratory syncytial virus
RSV2	C0086943	rous sarcoma virus
SLE1	C0014060	saint louis encephalitis
SLE2	C0024141	systemic lupus erythematosus
TPN1	C0027303	triphosphopyridine nucleotide
TPN2	C0030548	total parenteral nutrition
VCR1	C0042679	vincristine
VCR2	C0182936	videocassette recorder
VCR3	C0526312	vanadyl ribonucleoside complex

Appendix B. Detailed sense definitions for Set B

Sense	CUI	Semantic Categories
ASSOCIATION1	C0004083	Mental Process
ASSOCIATION2	C0699792	Social Behavior
COLD1	C0009264	Natural Phenomenon or Process
COLD2	C0009443	Disease or Syndrome
COLD3	C0024117	Disease or Syndrome
COLD4	C0010412	Therapeutic or Preventive Procedure
COLD5	C0234192	Qualitative Concept + Sign or Symptom
CULTURE1	C0010453	Idea or Concept
CULTURE2	C0430400	Laboratory Procedure
DEGREE1	C0449286	Qualitative Concept
DEGREE2	C0542560	Intellectual Product
DEPRESSION1	C0011570	Mental or Behavioral Dysfunction
DEPRESSION2	C0460137	Functional Concept
DISCHARGE1	C0012621	Body Substance
DISCHARGE2	C0030685	Health Care Activity
ENERGY1	C0424589	Finding
ENERGY2	C0542479	Natural Phenomenon or Process
EXTRACTION1	C0684295	Laboratory Procedure
EXTRACTION2	C0185115	Therapeutic or Preventive Procedure
FAT1	C0424612	Organism Attribute
FAT2	C0015677	Lipid
FIT1	C0036572	Sign or Symptom
FIT2	C0424576	Finding
FLUID1	C0302908	Substance
FLUID2	C0444611	Qualitative Concept
FREQUENCY1	C0439603	Temporal Concept
FREQUENCY2	C0042023	Sign or Symptom
GANGLION1	C0085648	Acquired Abnormality + Neoplastic Process
GANGLION2	C0017067	Body Part, Organ, or Organ Component
GLUCOSE1	C0017725	Carbohydrate + Biologically Active Substance
GLUCOSE2	C0337438	Laboratory Procedure
GROWTH1	C0018270	Organism Function
GROWTH2	C0220844	Functional Concept
IMPLANTATION1	C0029976	Organism Function
IMPLANTATION2	C0021107	Therapeutic or Preventive Procedure
INHIBITION1	C0021467	Mental Process
INHIBITION2	C0021469	Molecular Function
JAPANESE1	C0376247	Language
JAPANESE2	C0022342	Population Group

Sense	CUI	Semantic Categories
LEAD1	C0023175	Element, Ion, or Isotope
LEAD2	C0373667	Laboratory Procedure
MAN1	C0024554	Organism Attribute
MAN2	C0025266	Population Group
MAN3	C0086418	Human + Population Group
MOLE1	C0439189	Quantitative Concept
MOLE2	C0026386	Mammal
MOLE3	C0349514	Neoplastic Process
NUTRITION1	C0392209	Organism Attribute
NUTRITION2	C0028707	Organism Function + Biomedical Occupation or Discipline
NUTRITION3	C0600072	Therapeutic or Preventive Procedure
PATHOLOGY1	C0919386	Therapeutic or Preventive Procedure
PATHOLOGY2	C0677042	Pathologic Function
PRESSURE1	C0033095	Quantitative Concept
PRESSURE2	C0460139	Therapeutic or Preventive Procedure
PRESSURE3	C0234222	Organ or Tissue Function
REDUCTION1	C0441610	Health Care Activity
REDUCTION2	C0301630	Natural Phenomenon or Process
REPAIR1	C0374711	Therapeutic or Preventive Procedure
REPAIR2	C0043240	Organism Function
RESISTANCE1	C0683598	Social Behavior
RESISTANCE2	C0237834	Mental Process
SCALE1	C0222045	Body Part, Organ, or Organ Component
SCALE2	C0349674	Intellectual Product
SCALE3	C0175659	Manufactured Object
SECRETION1	C0036537	Body Substance
SECRETION2	C0036536	Biologic Function
SEX1	C0009253	Organism Function + Individual Behavior
SEX2	C0079399	Organism Attribute
SINGLE1	C0087136	Population Group
SINGLE2	C0205171	Quantitative Concept
STRAINS1	C0080194	Injury or Poisoning
STRAINS2	C0456178	Intellectual Product
SURGERY1	C0038894	Biomedical Occupation or Discipline
SURGERY2	C0038895	Functional Concept
TRANSIENT1	C0205374	Temporal Concept
TRANSIENT2	C0040704	Population Group
TRANSPORT1	C0005528	Cell Function
TRANSPORT2	C0150390	Health Care Activity
ULTRASOUND1	C0041618	Diagnostic Procedure
ULTRASOUND2	C0041621	Natural Phenomenon or Process
WEIGHT1	C0043100	Quantitative Concept
WEIGHT2	C0005910	Organism Attribute + Quantitative Concept
WHITE1	C0220938	Qualitative Concept
WHITE2	C0007457	Population Group

Appendix C. The detail corpus information for Set A

SID	GSS	STC_1	STC2	GSS-STC_2	STC_2∩GSS
ACE1	30	5,388	122	24	6
ACE2	5,820	3,192	7,288	1,240	4,580
ACE3	2	1	26	-	-
ACE4	2	-	21	1	1
ACE5	-	-	2	-	-
ACE6	1	-	1	1	-
ACE7	1	2	-	1	-
ANA1	53	81	9	48	5
ANA2	843	145	1,086	191	652
ANA3	-	28	6	-	-
APC1	1,356	1,817	3,445	21	1,335
APC2	430	2,258	550	33	397
APC3	8	578	15	2	6
APC4	37	1,175	4	3	34
APC5	479	854	144	365	114
ASP1	54	1,059	59	3	51
ASP2	17	965	29	12	5
ASP3	8	4,656	105	-	-
ASP4	2	2,298	14	1	1
ASP5	-	26	2	-	-
ASP6	60	4,820	96	4	56
BPD1	208	943	122	118	90
BPD2	465	917	313	210	255
BPD3	233	442	45	211	22
BPD4	-	-	14	-	-
BSA1	354	3,249	228	257	97
BSA2	2,808	3,445	1,180	2,589	219
CAD1	3,294	12,148	3,469	463	2,831
CAD2	-	-	6	-	-
CAD3	-	-	1	-	-
CAD4	16	34	12	4	12
CAD5	15	-	13	12	3
CAT1	34	2,569	887	21	13
CAT2	1	16,383	678	-	1
CAT3	-	156	2		
CAT4	1	13	1		1
CAT5	-	-	16	-	-
CML1	3,178	3,731	3,146	1,119	2,059
CML2	172	19	23	149	23

SID	GSS	STC_1	STC2	GSS-STC_2	$STC_2 \cap GSS$
CMV1	4,887	11	4,372	1,779	3,108
CMV2	4	34	5		4
CMV3	2	-	81	1	1
CMV4	51	145	112	23	28
CPI1	9	65	3	8	1
CPI2	59	75	27	45	14
CPI3	4	3	2	3	1
CSF1	9,961	4,023	7,176	5,772	4,189
CSF2	765	3,038	14,275	289	476
CSF3	44	72	45	36	8
CSF4	-	-	1	-	-
CVA1	226	10,445	399	54	172
CVA2	-	44	3	-	-
CVP1	6	1	63	3	3
CVP2	-	1	16	-	-
CVP3	581	12	132	517	64
DIP1	-	-			
DIP2	31	69	27	17	14
DIP3	81	-	40	61	20
DOB1	1	-	10		1
DOB2	-	-			-
DOB3	1	535	1		1
DVT1	1,584	1,695	472	323	1,261
DVT2	14	50	1,035	1	13
EMG1	-	68	31	-	-
EMG2	808	3,143	1,839	515	293
EMG3	2,036	37	407	1,725	311
EMG4	926	-		678	248
FDP1	8	-	155	7	1
FDP2	2	2		2	-
FDP3	382	454	552	164	218
FDP4	39	220	20	21	18
FDP5	-	-		-	-
HSV1	3,398	5,797	6,437	2,083	1,315
HSV2	81	542	57	40	41
IBD1	1,149	3,916	1,130	185	964
IBD2	-	1,470	61	-	-
LAM1	103	104	57	66	37
LAM2	-	4,372		-	-
LAM3	22	98	8	14	8
LAM4	2	-	40	2	-
LAM5	56	128	41	37	19

SID	GSS	STC_1	STC2	GSS-STC_2	$STC_2 \cap GSS$
LDH1	3,389	88	4,032	1,750	1,639
LDH2	1	3		1	-
MAC1	231	694	253	13	218
MAC2	40	8,506	430	3	37
MAC3	535	629	653	153	382
MAC4	-	-	31	-	-
MAC5	-	106	1	-	-
MAC6	18	20	13	18	-
MAC7	-	14		-	-
MAC8	19	-	7	12	7
MAC9	19	6	1	19	-
MAS1	31	121	20	13	18
MAS2	81	419	90	30	51
MAS3	-	24		-	-
MCP1	-	-	2	-	-
MCP2	8	-	49	1	7
MCP3	9	-		9	-
MCP4	157	2,450	109	82	75
MCP5	185	20	1,655	1	184
MCP6	102	180	189	29	73
PCA1	210	20	79	173	37
PCA2	376	348	88	341	35
PCA3	507	69	312	304	203
PCA4	5	357	40	3	2
PCA5	112	852	87	43	69
PCA6	-	-	-	-	-
PCA7	-	309	-	-	-
PCA8	343	1,704	70	315	28
PCA9	-	-	-	-	-
PCP1	-	18		-	-
PCP2	1	242	32		1
PCP3	341	352	122	245	96
PCP4	1,071	14	578	638	433
PCP5	812	1,836	1,004	151	661
PEG1	52	4,192	1,169	33	19
PEG2	18	325	51	13	5
PSA1	9	989	33	1	8
PSA2	3,215	604	1,345	2,310	905
PSA3	3	3	18		
PVC1	473	3,544	162	405	68
PVC2	98	3,564	261	12	86
PVC3	-	1	15		

SID	**GSS**	**STC_1**	**STC2**	**GSS-STC_2**	**STC_2∩GSS**
RSV1	1,335	1,190	457	1,033	302
RSV2	619	1,482	299	576	43
SLE1	138	14	107	57	81
SLE2	6,634	2,330	5,568	2,644	3,990
TPN1	2	5,645	46	1	1
TPN2	1,621	1,956	1,087	826	795
VCR1	634	4,586	606	215	419
VCR2	4	7	14	4	-
VCR3	-	5	1	-	-

Appendix D. The Detailed Corpus Information for Set B

SENSE	STC_2	STC_1
ASSOCIATION1	821	4
ASSOCIATION2	646	0
COLD1	288	373
COLD2	175	292
COLD3	280	840
COLD4	92	1
COLD5	128	0
CULTURE1	155	0
CULTURE2	154	127
DEGREE1	501	9
DEGREE2	13	31
DEPRESSION1	1291	420
DEPRESSION2	38	0
DISCHARGE1	156	0
DISCHARGE2	724	240
ENERGY1	273	200
ENERGY2	283	21
EXTRACTION1	76	0
EXTRACTION2	551	0
FAT1	388	1
FAT2	517	3
FIT1	240	330
FIT2	75	14
FLUID1	138	207
FLUID2	1123	11
FREQUENCY1	526	0
FREQUENCY2	81	490
GANGLION1	38	380
GANGLION2	1688	29
GLUCOSE1	1047	218
GLUCOSE2	93	124
GROWTH1	258	240
GROWTH2	0	0
IMPLANTATION1	208	608
IMPLANTATION2	557	314
INHIBITION1	624	5
JAPANESE1	11	36
JAPANESE2	172	200
LEAD1	167	6
LEAD2	0	0

SENSE	**STC_2**	**STC_1**
MAN1	248	632
MAN2	948	1
MAN3	4	206
MOLE1	19	0
MOLE2	61	0
MOLE3	59	124
NUTRITION1	1387	460
NUTRITION2	13	23
NUTRITION3	12	0
PATHOLOGY1	147	1
PATHOLOGY2	548	3
PRESSURE1	972	23
PRESSURE2	163	8
PRESSURE3	0	0
REDUCTION1	222	0
REDUCTION2	111	1
REPAIR1	548	1
REPAIR2	88	241
RESISTANCE1	71	17
RESISTANCE2	889	0
SCALE2	14	3
SCALE3	149	0
SECRETION1	467	10
SECRETION2	134	247
SEX1	91	681
SEX2	839	409
SINGLE1	5	0
SINGLE2	531	400
STRAINS1	4	93
STRAINS2	57	0
SURGERY1	171	9
SURGERY2	922	399
TRANSIENT1	199	200
TRANSPORT1	587	33
TRANSPORT2	1413	90
ULTRASOUND1	949	525
ULTRASOUND2	47	176
WEIGHT1	637	0
WEIGHT2	1006	406
WHITE1	217	56
WHITE2	218	468

Appendix E. The Detailed Semantic Relations Between Sense Definitions of Set A

AW	Direct Relatives	#(Relatives CUI)	The shortest Semantic distance (< 3)
ACE	ACE2(RO)ACE3	(ACE3,ACE5)1 (ACE4,ACE5)1 (ACE3,ACE4)2 (ACE1,ACE2)1	(ACE1,ACE2)0 (ACE2,ACE6)2 (ACE3,ACE4,ACE5,ACE7)0
ANA		(ANA2,ANA3)1	(ANA2,ANA3)0
APC			(APC2,APC3)1 (APC4,APC5)1
ASP	ASP3(CHD,RB,RN,RO)ASP6 ASP2(RO)ASP6	(ASP3,ASP6)21	(ASP2,ASP3,ASP5,ASP6)0 (ASP1,ASP4)2
BPD			(BPD1,BPD2)1 (BPD3,BPD4)0
CAD			(CAD1,CAD4)0 (CAD2,CAD3)0
CAT			(CAT1,CAT5)0
CMV			(CMV2,CMV3,CMV4)0
CSF			(CAF2,CSF3,CSF4)0
CVP		(CVP1,CVP2)6	(CVP1,CVP2)0
DOB		(DOB1,DOB3)1	
DVT	DVT1(RO)DVT2	(DVT1,DVT2)14	
EMG	EMG3(RO)EMG2 EMG4(RO)EMG2 EMG3(RO)EMG4	(EMG2,EMG3)2	(EMG2,EMG4)2
FDP		(FDP1,FDP2)2	(FDP1,FDP2,FDP3)0
IBD	IBD1(RO)IBD2	(IBD1,IBD2)24	
LAM	LAM3(RN)LAM5		(LAM3,LAM5)0 (LAM1,LAM4)2
LDH	LDH1(RO)LDH2		
MAC	MAC6(SIB)MAC9	(MAC6,MAC9)62	(MAC6,MAC9)0 (MAC4,MAC8)0
MAS			(MAS1,MAS2)0
MCP	MCP1(RO)MCP6		(MCP1,MCP3,MCP4)0 (MCP5,MCP6)0 ((MCP1,MCP3,MCP4), (MCP5,MCP6))1
PCA	PCA4(SIB)PCA5	(PCA4,PCA5)8	(PCA4,PCA5)0 (PCA3,PCA7)0 (PCA2,(PCA3,PCA7))2
PCP		(PCP2,PCP3,PCP4)1	(PCP2,PCP3,PCP4,PCP5)0*
RSV			(RSV1,RSV2)0
SLE			(SLE1,SLE2)0
VCR			(VCR1,VCR3)1

Appendix F. The performance for the best classifier for each combination of abbreviations and noise levels

AW	NL	P(%)	ML	AW	NL	P(%)	ML
ACE	0	99.3	MYDLL	FDP	0	99.4	MYDLL
ACE	0.05	99.1	TDLL	FDP	0.05	92.6	TDLL
ACE	0.1	99.1	TDLL	FDP	0.1	92.6	TDLL
ACE	0.15	98.6	TDLL	FDP	0.15	92.6	TDLL
ACE	0.2	98.2	TDLL	FDP	0.2	89.8	TDLL
ACE	0.25	96.9	TDLL	FDP	0.25	89.2	TDLL
ACE	0.3	94.5	TDLL	FDP	0.3	83.0	MSL,TDLL
ACE	0.35	90.2	TDLL	FDP	0.35	83.0	MSL
ACE	0.4	86.1	TDLL	FDP	0.4	83.0	MSL
ANA	0	100.0	MSL,NBL,MYDLL	HSV	0	99.9	MYDLL
ANA	0.05	100.0	MSL	HSV	0.05	98.6	TDLL
ANA	0.1	100.0	MSL	HSV	0.1	98.6	TDLL
ANA	0.15	100.0	MSL	HSV	0.15	98.5	MSL
ANA	0.2	100.0	MSL	HSV	0.2	98.5	MSL
ANA	0.25	100.0	MSL	HSV	0.25	98.5	MSL
ANA	0.3	100.0	MSL	HSV	0.3	98.5	MSL
ANA	0.35	100.0	MSL	HSV	0.35	98.5	MSL
ANA	0.4	100.0	MSL	HSV	0.4	98.5	MSL
APC	0	98.8	MSL	IBD	0	100.0	MSL,NBL,TDLL,MYDLL
APC	0.05	98.9	MSL	IBD	0.05	100.0	MSL,NBL,TDLL,MYDLL
APC	0.1	98.9	MSL	IBD	0.1	100.0	MSL,NBL,TDLL,MYDLL
APC	0.15	98.9	MSL	IBD	0.15	100.0	MSL,NBL,TDLL,MYDLL
APC	0.2	98.9	MSL	IBD	0.2	100.0	MSL,NBL,TDLL,MYDLL
APC	0.25	98.9	MSL	IBD	0.25	100.0	MSL,NBL,TDLL,MYDLL
APC	0.3	98.9	MSL	IBD	0.3	100.0	MSL,NBL,TDLL,MYDLL
APC	0.35	98.9	MSL	IBD	0.35	100.0	MSL,NBL,TDLL,MYDLL
APC	0.4	98.9	MSL	IBD	0.4	100.0	MSL,NBL,TDLL,MYDLL
ASP	0	86.7	TDLL	LAM	0	82.5	NBL
ASP	0.05	83.3	TDLL	LAM	0.05	80.0	MSL
ASP	0.1	78.3	MSL,TDLL	LAM	0.1	82.5	MSL
ASP	0.15	76.7	MSL	LAM	0.15	82.5	MSL
ASP	0.2	76.7	MSL	LAM	0.2	82.5	MSL
ASP	0.25	73.3	MSL	LAM	0.25	82.5	MSL
ASP	0.3	76.7	MSL	LAM	0.3	82.5	MSL
ASP	0.35	76.7	MSL	LAM	0.35	82.5	MSL
ASP	0.4	76.7	MSL	LAM	0.4	82.5	MSL
BPD	0	99.2	NBL	LDH	0	100.0	MSL
BPD	0.05	98.4	MSL	LDH	0.05	99.9	TDLL
BPD	0.1	98.4	MSL	LDH	0.1	99.7	TDLL
BPD	0.15	98.4	MSL	LDH	0.15	99.3	TDLL
BPD	0.2	98.4	MSL	LDH	0.2	98.3	TDLL
BPD	0.25	98.4	MSL	LDH	0.25	95.5	TDLL
BPD	0.3	98.4	MSL	LDH	0.3	91.3	TDLL
BPD	0.35	98.4	MSL	LDH	0.35	85.6	TDLL
BPD	0.4	98.4	MSL	LDH	0.4	82.6	MSL

AW	NL	P(%)	ML	AW	NL	P(%)	ML
BSA	0	98.0	MYDLL	MAC	0	95.5	TDLL,MYDLL
BSA	0.05	95.8	MSL	MAC	0.05	92.6	TDLL
BSA	0.1	95.8	MSL	MAC	0.1	92.0	TDLL
BSA	0.15	95.8	MSL	MAC	0.15	90.9	MSL
BSA	0.2	95.8	MSL	MAC	0.2	90.9	MSL
BSA	0.25	95.8	MSL	MAC	0.25	90.9	MSL
BSA	0.3	95.8	MSL	MAC	0.3	90.9	MSL
BSA	0.35	95.8	MSL	MAC	0.35	81.5	MSL
BSA	0.4	95.8	MSL	MAC	0.4	79.8	MSL
CAD	0	99.7	MYDLL	MAS	0	100.0	MSL,NBL,TDLL,MYDLL
CAD	0.05	99.2	TDLL	MAS	0.05	100.0	MSL
CAD	0.1	99.0	TDLL	MAS	0.1	100.0	MSL
CAD	0.15	99.0	TDLL	MAS	0.15	100.0	MSL
CAD	0.2	98.4	TDLL	MAS	0.2	100.0	MSL
CAD	0.25	96.8	TDLL	MAS	0.25	100.0	MSL
CAD	0.3	93.1	TDLL	MAS	0.3	100.0	MSL
CAD	0.35	88.8	TDLL	MAS	0.35	100.0	MSL
CAD	0.4	82.5	TDLL	MAS	0.4	100.0	MSL
CAT	0	95.0	MSL,TDLL,MYDLL	MCP	0	97.9	TDLL,MYDLL
CAT	0.05	95.0	MSL,TDLL	MCP	0.05	97.3	MSL
CAT	0.1	95.0	MSL	MCP	0.1	95.2	MSL
CAT	0.15	95.0	MSL	MCP	0.15	97.3	MSL
CAT	0.2	95.0	MSL	MCP	0.2	97.3	MSL
CAT	0.25	95.0	MSL	MCP	0.25	97.3	MSL
CAT	0.3	95.0	MSL	MCP	0.3	97.3	MSL
CAT	0.35	95.0	MSL	MCP	0.35	97.3	MSL
CAT	0.4	95.0	MSL	MCP	0.4	97.3	MSL
CML	0	98.4	MYDLL	PCA	0	97.9	MSL
CML	0.05	96.6	TDLL	PCA	0.05	93.3	MSL
CML	0.1	96.2	TDLL	PCA	0.1	91.6	MSL
CML	0.15	95.7	TDLL	PCA	0.15	91.6	MSL
CML	0.2	93.7	TDLL	PCA	0.2	91.6	MSL
CML	0.25	93.6	MSL	PCA	0.25	91.6	MSL
CML	0.3	93.6	MSL	PCA	0.3	91.6	MSL
CML	0.35	93.6	MSL	PCA	0.35	91.6	MSL
CML	0.4	93.6	MSL	PCA	0.4	91.6	MSL
CMV	0	99.7	MYDLL	PCP	0	96.8	MSL,MYDLL
CMV	0.05	99.3	TDLL	PCP	0.05	94.2	TDLL
CMV	0.1	99.4	TDLL	PCP	0.1	92.1	MSL,TDLL
CMV	0.15	99.1	TDLL	PCP	0.15	92.1	MSL
CMV	0.2	98.7	TDLL	PCP	0.2	92.1	MSL
CMV	0.25	97.5	TDLL	PCP	0.25	92.1	MSL
CMV	0.3	94.1	TDLL	PCP	0.3	92.1	MSL
CMV	0.35	90.4	TDLL	PCP	0.35	92.1	MSL
CMV	0.4	85.2	MSL	PCP	0.4	92.1	MSL

AW	NL	P(%)	ML	AW	NL	P(%)	ML
CPI	0	97.2	TDLL,MYDLL	PEG	0	100.0	MSL
CPI	0.05	97.2	MSL	PEG	0.05	96.9	MSL
CPI	0.1	97.2	MSL	PEG	0.1	96.9	MSL
CPI	0.15	97.2	MSL	PEG	0.15	96.9	MSL
CPI	0.2	97.2	MSL	PEG	0.2	96.9	MSL
CPI	0.25	97.2	MSL	PEG	0.25	96.9	MSL
CPI	0.3	97.2	MSL	PEG	0.3	96.9	MSL
CPI	0.35	97.2	MSL	PEG	0.35	96.9	MSL
CPI	0.4	97.2	MSL	PEG	0.4	96.9	MSL
CSF	0	98.2	MYDLL	PSA	0	99.8	TDLL
CSF	0.05	95.8	TDLL	PSA	0.05	99.8	TDLL
CSF	0.1	95.5	TDLL	PSA	0.1	99.8	TDLL
CSF	0.15	94.8	TDLL	PSA	0.15	99.2	TDLL
CSF	0.2	93.8	TDLL	PSA	0.2	98.5	TDLL
CSF	0.25	91.2	TDLL	PSA	0.25	97.5	MSL,TDLL
CSF	0.3	90.2	MSL	PSA	0.3	97.5	MSL
CSF	0.35	90.2	MSL	PSA	0.35	97.5	MSL
CSF	0.4	90.2	MSL	PSA	0.4	97.5	MSL
CVA	0	100.0	MSL,NBL,TDLL,MYDLL	PVC	0	97.8	MSL,NBL,MYDLL
CVA	0.05	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.05	97.8	MSL
CVA	0.1	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.1	97.8	MSL
CVA	0.15	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.15	97.8	MSL
CVA	0.2	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.2	97.8	MSL
CVA	0.25	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.25	97.8	MSL
CVA	0.3	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.3	97.8	MSL
CVA	0.35	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.35	97.8	MSL
CVA	0.4	100.0	MSL,NBL,TDLL,MYDLL	PVC	0.4	97.8	MSL
CVP	0	99.6	TDLL,MYDLL	RSV	0	97.5	TDLL,MYDLL
CVP	0.05	99.2	MSL,TDLL	RSV	0.05	97.1	MSL
CVP	0.1	99.2	MSL	RSV	0.1	97.1	MSL
CVP	0.15	99.2	MSL	RSV	0.15	97.1	MSL
CVP	0.2	99.2	MSL	RSV	0.2	97.1	MSL
CVP	0.25	99.2	MSL	RSV	0.25	97.1	MSL
CVP	0.3	99.2	MSL	RSV	0.3	97.1	MSL
CVP	0.35	99.2	MSL	RSV	0.35	97.1	MSL
CVP	0.4	99.2	MSL	RSV	0.4	97.1	MSL
DIP	0	100.0	MSL,NBL,TDLL,MYDLL	SLE	0	100.0	MYDLL
DIP	0.05	100.0	MSL	SLE	0.05	98.9	TDLL
DIP	0.1	100.0	MSL	SLE	0.1	99.0	TDLL
DIP	0.15	100.0	MSL	SLE	0.15	98.8	TDLL
DIP	0.2	100.0	MSL	SLE	0.2	97.7	TDLL
DIP	0.25	100.0	MSL	SLE	0.25	95.2	TDLL
DIP	0.3	100.0	MSL	SLE	0.3	91.4	TDLL
DIP	0.35	100.0	MSL	SLE	0.35	91.0	MSL
DIP	0.4	100.0	MSL	SLE	0.4	91.0	MSL

AW	NL	P(%)	ML	AW	NL	P(%)	ML
DOB	0	0.0	*	TPN	0	100.0	MSL,TDLL
DOB	0.05	12.5	NBL,TDLL,MYDLL	TPN	0.05	100.0	TDLL
DOB	0.1	25.0	NBL,TDLL,MYDLL	TPN	0.1	100.0	TDLL
DOB	0.12	0.0	*	TPN	0.15	99.7	TDLL
DOB	0.2	25.0	NBL,TDLL,MYDLL	TPN	0.2	98.5	TDLL
DOB	0.25	25.0	NBL,TDLL,MYDLL	TPN	0.25	96.3	TDLL
DOB	0.3	25.0	NBL,TDLL,MYDLL	TPN	0.3	93.4	TDLL
DOB	0.35	37.5	NBL,TDLL,MYDLL	TPN	0.35	86.1	TDLL
DOB	0.4	37.5	NBL,TDLL,MYDLL	TPN	0.4	82.8	MSL
DVT	0	99.4	TDLL,MYDLL	TPN	0.4	82.8	MSL
DVT	0.05	99.4	TDLL	VCR	0	100.0	MYDLL
DVT	0.1	99.2	TDLL	VCR	0.05	99.2	MSL,TDLL
DVT	0.15	98.9	TDLL	VCR	0.1	99.2	MSL
DVT	0.2	98.3	TDLL	VCR	0.15	99.2	MSL
DVT	0.25	95.7	TDLL	VCR	0.2	99.2	MSL
DVT	0.3	92.5	MSL	VCR	0.25	99.2	MSL
DVT	0.35	92.5	MSL	VCR	0.3	99.2	MSL
DVT	0.4	92.5	MSL	VCR	0.35	99.2	MSL
EMG	0	58.3	TDLL	VCR	0.4	99.2	MSL
EMG	0.05	56.9	TDLL				
EMG	0.1	54.4	TDLL				
EMG	0.15	52.3	TDLL				
EMG	0.2	51.6	MSL				
EMG	0.25	51.6	MSL				
EMG	0.3	51.6	MSL				
EMG	0.35	51.6	MSL				
EMG	0.4	51.6	MSL				

Reference List

[1] UMLS Knowledge Sources, US Dept of Health and Human Services, National Institutes of Health, National Library of Medicine.

[2] Agirre E, Martinez D *Exploring automatic word sense disambiguation with decision lists and the Web*. COLING 2000.

[3] Agirre E, Rigau G. *A proposal for word sense disambiguation using conceptual distance*. Amsterdam Studies in the Theory and History of Linguistic Sicence: Current Issues in Linguistic Theory. 1997: 161-173.

[4] Agirre E, Rigau G *Word Sense Disambiguation using Conceptual Density*. COLING 1996.

[5] Aha D, Kibler D, Albert M. *Instance-based learning algorithms*. Machine Learning 1991; 7:33-66.

[6] Aronson AR, Rindflesch TC, Browne A *Exploiting a large thesaurus for information retrieval*. 1994; Proc RIAO' 94: 197-216.

[7] Aronson A *Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program*. Proc. AMIA Symp 2001; 2001: 17-21.

[8] Bloom D. *Acronyms, abbreviations and initialism*. BJU International 2000; 86:1-6.

[9] Brill E. *Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging.* Computational Lingustics 21, 543-565.

[10] Brown PF, Della Pietra VJ, DeSouza P, Lai J, Mercer RL. *Class-based n-gram models of natural language*. Computational Lingustics 1992; 18(4):467-479.

[11] Bruce R, Wiebe J *Word-Sense Disambiguation Using Decomposable Models*. Proceedings of the Thirty-Second Annual Meeting of The Association from Computational lingustics 1994: 139-146.

[12] Bruce R, Wiebe J *Word Sense distinguishability and inter-coder agreement*. Proc. of ACL 1998; 32: 139-145.

[13] Campbell D, Johnson SB *Comparing Syntactic Complexity in Medical and non-Medical Corpora*. Proc. AMIA 2001 Symp.: 90-94.

[14] Cardie C. dissertation *Domain Specific Knowledge Acquisition for Conceptual Sentence Analysis*. 1994.

[15] Carletta J. *Assessing agreement on classification tasks: the kappa statistics*. Computational Linguistics 1996; 22(2):249-254.

[16] Chapman R. Roget's International Thesaurus. 1977. Harper and Row, New York.

[17] Chen N, Chang J. *Topical Clustering of MRD Senses Based on Information Retrieval Techniques*. Computational Linguistics 1998; 24(1):61-95.

[18] Cheung T. *Acronyms in clinical traials in cardiology --1998*. Am Heart J 1999; 134(4(1)):726-765.

[19] Chi J, Seo J, Kim G. *Verb sense ddisambiguation based on dual distributional similarity*. Nat Lang Eng 1999; 5(2):157-170.

[20] Church KW *A stochastic parts program and noun phrase parser for unrestricted text*. Second Conference on Applied Natural Language Processing 1988: 136-143.

[21] Church K, Mercer R. *Introduction to the special issue on conputational linguistics using lrge corpora*. Computational Lingustics 1993; 19(1):1-24.

[22] Cohen J. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement 1960; 20:37-46.

[23] Cowie J, Guthrie J, Guthrie L *lexical disambiguation using simulated annealing*. Proceedings of COLING Conference 1992: 359-365.

[24] Dagan I, Lee L, Pereira F *Similarity-based methods for word sense disambiguation*. Proceedings of 35th ACL-EACL 1997: 56-63.

[25] Dagan I, Lee L, Pereira F *Similarity-based Models of Word Cooccurrence Probabilities*. Machine Learning 1999

[26] Duda R, Hart P. *Pattern Classification and Scene Analysis*. John Wiley and Sons, NY, 1973.

[27] Engelson SP, Dagan I *Minimizing manual annotation cost in supervised training from corpora*. 1996; ACL 34: 319-326.

[28] Escudero G, Marquez and Rigau G *Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited*. Proc. 14th ECAI 2000

[29] Fellbaum C. *WordNet: An Electronic Lexical Database*. 1998.

[30] Fellbaum C, Crabowski J, Landes S, Baumann A. *Matching words to senses in WordNet: Naive vs. expert differentiation of senses*. In: Fellbaum C, editor.

WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press, 1996.

[31] Friedman C *A Broad Coverage Natural Language Processing System*. Proc AMIA Symp 2000: 270-274.

[32] Friedman C, Hripcsak G. *Evaluating natural language processors in the clinical domain*. Meth Inf in Med 1998; 37:334-344.

[33] Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. *Natural language processing in an operational clinical information system*. Nat Lang Eng 1995; 1(1):83-108.

[34] Friedman C, Liu H, Shagina L, Johnson SB, Hripcsak G *Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing*. 2001; Proc AMIA Symp 2001: 189-193.

[35] Fujii A, Inui K, Tokunaga T, Tanaka H. Selective sampling for example-based word sense disambiguation. Computational Lingustics 24[4], 573-597. 1998.

[36] Gale WA, Church KW, Yarowsky D *One Sense Per Discourse*. Proceedings of the ARPA Workshop on Speech and Natural Language Processing 1992: 233-237.

[37] Gale WA, Church KW, Yarowsky D *Using bilingual materials to develop word sense disambiguation methods*. 1992; Proceedings of the 4th International Conference on Theoretical and Methodlogical Issues in Machine Translation: 101-112.

[38] Hanks P. *Collins English Dictionary*. 1979.

[39] Haykin S. *Neural Networks*. 1994.

[40] Hearst MA *Noun homograph disambiguation using local context in large text corpora*. 1991; Seventh Annual Conference of the UW Centre for the New OED and Text Research: 59-68.

[41] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. *Unlocking clinical data from narrative reports*. Ann of Int Med 1995; 122(9):681-688.

[42] http:, www.cpmc.columbia.edu/resources/cis/repository.html. NYPH Clinical Data Repository. [2001]. 2001.

[43] Humphrey S, Rindflesch TC, Aronson AR *Automatic indexing by discipline and high-level categories: methodology and potential applications*. Proc. 11th ASIST SIG/CR Classification Research Workshop 2000

[44] Ide N, Veronis J *Very large neural networks for word-sense disambiguation*. *9th European Conference on Artificial Intelligence* 1990: 366-368.

[45] Ide N, Veronis J. *Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art.* Computational Lingustics 1998; 24(1): 1-40.

[46] Johnson SB. *A semantic lexicon for medical language processing*. J Am Med Inf Assoc 1999; 6(3):205-218.

[47] Jorgensen J. *The psychological reality of word senses*. Journal of Psycholinguistic Research 1990; 19(3):167-190.

[48] Jurafsky D, Martin J. *Word Sense Disambiguation and Information Retrieval*. Speech and Language Processing. Prentice Hall, 2000: 631-666.

[49] Karov Y, Edelman S. *Similarity-based Word Sense Disambiguation.* Computational Lingustics 1998; 24(1):41-59.

[50] Kaufman L, Rousseeuw P. *Finding groups in data*. New York: Wiley, 1990.

[51] Kelly E, Stone P. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam, 1975.

[52] Kikui G *Term-list translation using monolingual co-occurrence vectors*. Proceedings of COLING-ACL 98 1998

[53] Kikui G *Resolving Translation Ambiguity using Non-parallel Bilingual Corpora*. ACL'99 Workshop on Unsupervised Learning in Natural Language Processing 1999

[54] Kilgarriff A. *Dictionary word sense distinctions: An enquiry into their nature*. Computers and the Humanities 1993; 26(1-2):365-387.

[55] Kilgarriff A. *I don't believe in word senses*. Computers and the Humanities 1997; 31(2):91-113.

[56] Kilgarriff A. *Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs*. Computer Speech and Language 1998; 12(3).

[57] Kilgarriff A, Rosenzweig J *Framework and results for English SENSEVAL*. 2000; Computers and the Humanities: 1-2.

[58] Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. *Respiratory isolation of tuberculosis patients using clinical guidelines and an automated decision support system*. Infection Control and Hospital Epidemiology 1998; 19(2):94-100.

[59] Krovetz R. dissertation *Word Sense Disambiguation for Large Text Databases*. 1995.

[60] Krovetz R *More than One Sense Per Discourse*. Proceedings of the ACL-SIGLEX Workshop 1998

[61] Krovetz R, Croft W. *Lexical ambiguity and information retrieval*. ACM Transactions on Information Systems 1992; 10(2):115-141.

[62] Kucera H, Francis W. *Computational analysis of present-day American English*. 1967.

[63] Larkey L, Ogilvie P, Price A, Tamilio B. *Acrophile: An Automated Acronym Extractor and Server*. ACM Digital Libraries 2000;205-214.

[64] Leacock C, Chodorow M, Miller G. *Using Corpus Statistics and WordNet relations for Sense Identification*. Computational Lingustics 1998; 24(1):147-165.

[65] Leacock C, Towell G, Voorhees EM *Corpus-based statistical sense resolution*. Proceedings of the ARPA Workshop on Human Language Technology; 1993.

[66] Lesk M *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an cie cream cone*. Proceeding of SIGDOC 1986: 24-26.

[67] Lewis DD, Gale WA *A sequentional algorithm for training text classifier*. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 3-12.

[68] Liu H, Lussier Y, Friedman C *A study of the UMLS abbreviations*. Proc. AMIA Symp. 2001: 393-397.

[69] Liu H, Lussier Y, Friedman C. *Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method*. Journal of Biomedical Informatics 2001; 34(4):249-261.

[70] Luk A *Statistical sense disambiguation with relatively small corpora using dictionary definitions*. Proceedings of the 33rd Annual Meeting of the ACL 1995; 33: 181-188.

[71] Luxton T, Al-Qassab H. *Better use of abreviations - a lesson from a stroke unit*. Medical Education; 2000(34):965-965.

[72] Marquez L. *Machine Learning and Natural Language Processing*. Machine Learning 2000.

[73] Martinez D, Agirre E *One sense per collocation and genre/topic variations*. 2000

[74] Masterman M. *The thesaurus in syntax and semantics*. Mechanical Translation 1957; 4(1):1-2.

[75] McRoy S. *Using multiple knowledge sources for word sense discrimination*. Computational Lingustics 1992; 18(1):1-30.

[76] Milhalcea R, Moldovan D *A Method for Word Sense Disambiguation of Unrestricted Text*. Proceedings of ACL 1999

[77] Milhalcea R, Moldovan D *An Automatic Method for Generating Sense Tagged Corpora*. AAAI 1999

[78] Milhalcea R, Moldovan D. *A highly accurate bootstrapping algorithm for word sense disambiguation*. International Journal of Artificial Intelligence Tools 2001; 10(1-2).

[79] Mitchell T. Machine Learning. McGraw Hill, 1997.

[80] Mooney R *Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning*. Proceedings of the Conference on EMNLP 1996: 82-91.

[81] Mooney R. Inductive Logic Programming for Natural Language Processing. Muggleton, editor. 1997. Springer Verlag. Inductive Logic Programming: Selected Papers from the 6th International Workshop.

[82] Nadkarni P, Chen R, Brandt C. *UMLS Concept Indexing for Production Databases*. J Am Med Inf Assoc 2001; 8:80-91.

[83] Ng HT *Examplar-Based Word Sense Disambiguation: Some Recent Improvements*. Proc EMNLP-2; 1997.

[84] Ng HT *Getting serious about word-sense disambiguation*. 1997; Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?: 1-7.

[85] Ng HT, Lee HB *Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach*. 1996; ACL 34: 40-47.

[86] Ng HT, Zelle J. *Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing*. AI Magazine 1997[Winter], 45-64.

[87] Pedersen T, Bruce R *Distinguishing Word Senses in Untagged Text*. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing 1997

[88] Pereira F, Tishby N, Lee L *Distributional clustering of English words*. Proceedings of the 31st Annual Meeting of the ACL 1993: 183-190.

[89] Porter MF. *An algorithm for suffix stripping.* Program 1980; 14(3):130-137.

[90] Proctor P. *Longman Dictionary of Contemporary English.* 1978. Longman Group, Essex, England.

[91] Quinlan J. *Induction of Decision Trees*. Machine Learning 1986; 1:81-106.

[92] Ramshaw L, Marcus M *Text chunking using transformation-based learning*. Proceedings of the Third Annual Workshop on Very Large Corpora 1995; 3: 82-94.

[93] Resnik P *WordNet and distributional analysis: A class-based approach to lexical discovery*. AAAI Workshop on statistically-based Natural Language Processing Techniques 1992

[94] Resnik P *Selectional Preference and Sense Disambiguation*. Proceedings of ASL/SIGLEX 1997

[95] Rindflesch TC, Aronson AR *Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus*. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care 1994: 240-244.

[96] Rindflesch TC, Hunter L, Aronson AR *Mining molecular binding terminology from biomedical text*. Proc AMIA Symp 1999: 127-131.

[97] Roth L, Hole W *Managing Name Ambiguity in the UMLS Metathesaurus*. Proc. AMIA Symposium 2000.

[98] Sanderson M. *Retrieving with good sense*. Information Retrieval; 2000.

[99] Schütze H *Dimensions of Meaning*. Proceedings of Supercomputing 1992; 92: 787-796.

[100] Schütze H. *Automatic Word Sense Discrimination.* Computational Linguistics 24(1): 97-123.

[101] Shortliffe EH, Wiederhold G, Perreault L, Fagan L. *Medical Informatics: Computer Applications in Health Care and Biomedicine*. Springer Verlag, 2000.

[102] Sneiderman CA, Rindflesch TC, Aronson AR *Finding the Findings: Identification of Findings in Medical Literature Using Restricted Natural Language Processing*. Proceedings of the 1996 AMIA Fall Symposium: 239-243.

[103] Spyns P. *Natural Language Processing in Medicine: An Overview*. Meth Inform Med 1996; 35:285-301.

[104] Stevenson M, Wilks Y *Combining weak knowledge sources for sense disambiguation*. 1999; Proc. of International Joint Conference on AI

[105] Sussna M *Word sense disambiguation for free-text indexing using a massive semantic network*. Proc. of the International Conference on Information and Knowledge management 1993: 67-74.

[106] Swanson DR. *Migraine and magnesium: Eleven neglected connections*. Perspect Biol Med 1988; 31(4):526-557.

[107] Taghva K, Gilbrech J. Recognizing Acronyms and their Definitions. 95-03. 1995. http://www.isri.unlv.edu/ir/publications/Taghva95-03.ps.

[108] Theodoridis S, Koutroumbas K. *Pattern Recognition*. Academic Press, 1998.

[109] Towell G, Voorhees EM. *Disambiguating Highly Ambiguous Words*. Computational Lingustics 1998; 24(1):125-146.

[110] Vapnik. Statistical Learning Theory. 1998. John Wiley and Sons.

[111] Veronis J *A study of polysemy judgements and inter-annotator agreement*. *Programme and advanced papers of the Senseval workshop* 1998

[112] Veronis J, Ide N *Very large neural networks for natural language processing*. *International Neural Network Conference* 1990

[113] Weeber M, Klein H, Aronson A, Mork J *Text-based discovery in biomedicine: the architecture of the DAD-system*. Proc AMIA Symp 2000: 903-907.

[114] Weeber M, Mork J, Aronson A *Developing a Test Collection for Biomedical Word Sense Disambiguation*. Proc. AMIA 2001 Symp: 746-750.

[115] Weiss S. *Learning To Disambiguate*. Information Storage and Retrieval 1973; 9:33-41.

[116] Wilks Y. *Senses and Texts*. Computers and the Humanities 1997.

[117] Wilks Y, Fass D, Guo CMJ, Plate T, Slator B. *A tractable machine dictionary as a resource for computational semantics*. Computational Lexicography for Natural Language Processing. 1989: 193-228.

[118] Wilks Y, Stevenson M *Word Sense Disambiguation using Optimized Combinations of Knowledge Sources*. 2000

[119] Witten I, Bell T. *The zero-frequency problem: Estimating the probabilities of novel events in asaptive text compression*. IEEE Transaction on Information Theory 1991; 37(4):1085-1094.

[120] Wolff S. *The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding*. Meth Inf in Med 1984; 23:195-203.

[121] Yarowsky D *Word sense disambiguation using statistical models of Roget's categories trained on large corpora*. 1992; Proceeding of the 14th International Conference on Computational Lingustics: 454-460.

[122] Yarowsky D *One sense per Collocation*. Proceeding of the 5th DARPA Speech and Natural Language Workshop 1993

[123] Yarowsky D *Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French*. 1994; ACL 32: 88-95.

[124] Yarowsky D *Unsupervised word sense disambiguation rivaling supervised methods*. 1995; ACL 33: 189-196.

[125] Yoshida M, Fukuda K, Takagi T. *PNAD-CSS: a workbench for construction a protein name abbreviation dictionary*. Bioinformatics 2000; 16(2):169-175.

[126] Yu H, Hripcsak G, Friedman C. *Mapping abbreviations to full forms in electronic articles*. JAMIA 2002(9):262-272.