Automatic word sense disambiguation of acronyms and abbreviations in clinical texts

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

Sungrim Moon

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Serguei V.S. Pakhomov, PhD (Advisor)

December 2012

Acknowledgements

I would like to express my sincere appreciation to my research mentor, Genevieve Melton-Meaux and my advisor, Dr. Serguei Pakhomov for all their guidance and support throughout this research. I would also like to acknowledge Dr. Stuart Speedie, Dr. Terrence Adam, and Dr. David Pieczkiewicz for their inspiration and advice serving as my committee members.

I would like to thank my parents Kwangpil Moon and Jungshin Lee, my sister Sungsook Moon and my brother Sungjoon Moon, and friends Yeonju Choi and Hyewon Lee for their encouragement and support. I also appreciate the support of my colleagues in NLP/IE: Yan Wang, Robert Bill, Rui Zhang, James Ryan, Bridget McInnes, Ying Liu, and Trevor Wennblom. Furthermore, I wish to thank all my colleagues at the Institute for Health Informatics, especially Youngtaek Park. Finally, I would like to thank Jessica Whitcomb-Trance and Mike Grove for their consultation about writing. Without the help of all the people above, I could not have finished this thesis.

Dedication

To my parents Kwangpil Moon and Jungshin Lee, my sister Sungsook Moon, and my brother Sungjoon Moon. Your support and encouragement for my whole life has made this work possible.

Abstract

The use of acronyms and abbreviations is increasing profoundly in the clinical domain in large part due to the greater adoption of electronic health record (EHR) systems and increased electronic documentation within healthcare. A single acronym or abbreviation may have multiple different meanings or senses. Comprehending the proper meaning of an acronym or abbreviation using automated machine techniques, known as word sense disambiguation (WSD), in clinical notes is an essential step for medical natural language processing (NLP) systems. While acronym and abbreviation WSD from the biomedical literature is an active area of investigation, little research has been done on this topic with clinical documents.

The purpose of this dissertation is to develop automatic WSD tools for clinical acronyms and abbreviations. A key step toward this end is to build a comprehensive clinical sense inventory based upon the integration of available biomedical resources and upon senses from a large corpus of clinical notes. Another complementary task is the performance maximization of machine learning (ML) algorithms. This includes the development of optimal feature sets for WSD and the exploration of minimum "adequate" sample size for training classifiers. These automatic WSD technologies extend to the complementary problem of symbol disambiguation in clinical texts. Lastly, the anticipated future work will be in developing quality improvement of automatic WSD tools including sense amelioration utilizing biomedical knowledge.

Table of Contents

List of Tables

List of Figures

Chapter 1 INTRODUCTION

1.1 Significance

Under time restraints that most clinicians experience in the clinical environment, compressed expressions, mainly acronyms and abbreviations, are widely utilized[1-4]. These compressed expressions are an efficient and conventional method of communication within clinical documents and are used daily extensively. In addition to traditional use, the need to understand and deal with these expressions is becoming more important because of the widespread adoption of electronic health record (EHR) systems and greater numbers of electronic clinical notes for documentation and communication in clinical care.

The widespread use of acronyms and abbreviations in clinical texts demands appropriate sense resolution among associated multiple meanings/senses for effective document utilization and for patient safety in clinical care[5, 6]. Correspondingly, automated resolution of the correct concept[7], or sense of an acronym or abbreviation, in the given text is considered a specialized type of word sense disambiguation (WSD)[8]. Accomplishing automatic medical natural language processing (NLP) requires acronym and abbreviation WSD automation to effectively utilize these documents for automated purposes[2, 4, 9]. While a human can properly comprehend what something means given approximately five words including an acronym or abbreviation in the center position, machine automation for interpreting specific senses within the document context continues to be a major challenge for automated medical NLP systems[7, 10].

One of the major inherent difficulties of clinical WSD problem is the informal

nature of clinical documentation and lack of formal structure to clinical notes[1, 3, 11]. For instance, clinical notes contain short telegraphic phrases and include structured reporting, resulting in a variety of data types within the clinical notes. Because of the powerful freedom of expression that the text provides, clinical notes continue to remain important for communication. These notes are, however, created in error-prone, time-constrained conditions with rare formalization. Typographical errors are also common due to (1) the lack of adequate spell checking/correction[11] and (2) misinterpretation between clinicians who may dictate meaning and the transcriptionists who must interpret what has been said.

In addition, because of the strict privacy issues of patient confidentiality, use of clinical documents for research is a substantial ongoing and extrinsic hurdle. Most clinical notes are produced outside of nationwide standards or agreements for document sharing. The Health Insurance Portability and Accountability Act (HIPAA), in particular, makes it difficult to access or share clinical documents for research not only from a single institution but also inter-institutionally[12].

Because of the challenges associated with the WSD problem including document access and the informal nature of clinical text, there is no comprehensive, open-source clinical sense inventory for clinical acronyms and abbreviations. Instead, there are a few small sets of unstandardized acronyms and abbreviations some of which contain sense inventories. Hence, to date, there has been minimal research about resources, techniques and tools, and other practical considerations for resolving ambiguous acronyms and abbreviations in clinical domains[11, 12].

1.2 Background

To perform WSD of clinical acronyms and abbreviations automatically, it is necessary to look at other overlapping multi-disciplinary fields including biomedical NLP and general English computational linguistics.

1.2.1 Available sense Inventories in biomedicine

The central basic assumption of biomedical sense inventories is that the short and long form of the acronym or abbreviation occurs closely with or without parentheses in a document[3]. For example, "comprehensive metabolic panel (CMP)" fulfills this assumption. One of the seminal works using this was the development of a character-mapping algorithm by Schwartz and Hearst[13]. Here, the authors identified pairs of SF and LF by matching characters between the short form and first characters of the long form. It is a simple heuristic algorithm but has very good performance. SaRAD[14], ARGH[15], ALICE[16] are also rule-based methods and databases building upon this using slightly different rules and approaches within the biomedical domain.

Machine learning (ML) algorithms, which will be discussed in section 1.2.3 in detail, have also been employed for biomedical acronym and abbreviation sense inventory creation. For example, Chang et al. created the Stanford biomedical abbreviation server[17]. It is based on using the longest common subsequence (LCS) algorithm[18] with a supervised logistic regression as a relevance evaluator for the short and long form pairs. From these techniques, they were able to discover 64,242 pairs of acronym and abbreviation SF and LF from 2004 MEDLINE abstracts. In the clinical

domain, Xu et al. focused on 12 clinical acronyms and abbreviations from a corpus of 16,949 admission notes to study the "Annotation Cost" and the "Sense Completeness" for building sense inventories[19]. The authors created a sense inventory using the Expectation Maximization clustering ML algorithm with the minimum manual annotation.

Statistical approaches can be also used to create sense inventories. In biomedicine, Liu and Friedman filtered irrelevant short and long forms using a collocation-based approach after detecting the parenthetical expressions in biomedical documents[20]. Collocation is the adjacent word collection near an item of interest. From this, 381,126 pairs of acronyms and abbreviations were identified. A well-known biomedical sense inventory, Another Database of Abbreviations in MEDLINE (ADAM)[21], is based on rules of length ratio and empirical cut-off values to filter out insignificant pairs. This approach was applied to titles and abstracts from 2006 MEDLINE, and 59,403 pairs of short and long forms were identified.

There are several additional biomedical acronym and abbreviation sets. The MEDLINE Abbreviation collection by Liu et al. examined 35 three character acronyms and abbreviations from abstracts of MEDLINE citations[22]. It utilizes the medical ontology, The Unified Medical Language System (UMLS)[23], by a supervised ML method. As a clinical sense set, one of few collections is the Mayo Clinic acronym and abbreviation set. This consists of physician-annotation of 16 acronyms and abbreviations resulting in 141 pairs based upon a subset of 17 million clinical notes.

1.2.2 Features used with machine learning and other automated approaches

To discuss different classes of potential features, it is necessary to examine related research in WSD within the fields of biomedical NLP, computational linguistics, statistics, and the clinical domains. These can be categorized as domain knowledge-based, linguistic, statistical, and general document features[24]. Unique or combined features are used for inputs in ML algorithms. Finding the optimal feature sets and consisdering the strengths and weaknesses of each individual feature type is a critical cornerstone to achieving high performance of ML techniques for these tasks[24, 25].

1.2.2.1 Domain knowledge-based features

The Unified Medical Language System (UMLS) and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[26], distributed by the National Library of Medicine (NLM) with the later maintained by IHSDO, are well-known medical terminology resources and extensively used by the medical NLP community. Medical domain knowledge-based features extracted from these resources are reasonable for use with clinical NLP tasks because clinical documents are based upon applications of medical knowledge.

Domain knowledge-based features are powerful since the UMLS offers not only term identification but also complementary information such as semantic relationships, semantic types, and semantic groups[7, 12, 27-29]. SNOMED CT may have better potential compared to many other resources since it is a clinical reference terminology[30]. Since concepts between the biomedical and clinical domain overlap[31], utilizing either resource

with automatic NLP tools is reasonable, as well. Therefore, automatic tools for the UMLS, such as the Specialist Lexicon[32] or MetaMap[33], can be readily implemented for term normalization, stemming, and semantic grouping in the clinical domain[33, 34].

On the other hand, use of available biomedical domain resources may increase the false positive or false negative rate when applying these resources to the clinical domain[9], particularly because terms in clinical terminologies are covered in variable ways in the UMLS[3, 12, 31]. One study of the 2009 UMLS version demonstrated low coverage of clinical terms, estimated at approximately 35%[3]. Moreover, the UMLS has noise, inconsistencies, and vague concepts because of the complicated and large accumulated biomedical knowledge that results in uneven hierarchical structures[9, 29, 34-36]. As a result, ongoing, additional curative work is essential[4, 31, 37].

Several representative studies have used domain knowledge-based features for WSD using unambiguous synonyms[38] or using the hierarchical relationship in the UMLS[22]. For instance, McInnes et al. utilized UMLS Concept Unique Identifiers (CUIs) for disambiguation with a supervised ML algorithm[39]. CUIs are the biomedical concepts of the UMLS. Leroy and Rindflesh used UMLS semantic types and relationships in the UMLS semantic network to apply supervised ML techniques for WSD tasks[28]. MetaMap can generate this semantic information and CUI automatically. However, semantic information as a feature for ML has limitations compared to CUI because of its lower granularity, and may have inconsistent results since any term may belong to multiple semantic types or have inconsistencies in the UMLS[28, 29].

Another type of medical resource, namely medical dictionaries, such as *Stedman's*

or *Dorland's*, can be useful resources because they define the term with associated semantic descriptions in a well-distinguished way with high accuracy. These resources have been historically less utilized for automatic medical NLP. One of the reasons is because of copyright issues. These resources also mix terms from the basic science and the clinical domain, which may result in additional non-applicable mappings or senses. Furthermore, while there can sometimes be rapid changes in clinical language terms, these resources may not always be updated, which can potentially be problematic.

While medical dictionaries for automated WSD have not been used to date in medical field, in general English, Lesk described an algorithm[40] using dictionary definitions for general English WSD using WordNet[41]. WordNet is available as an electronic open-source tool from the general English domain. This algorithm uses the definition of the target term compares words from the ambiguous target term in the text. This algorithm then assigns the maximum match as the assigned sense. An analogous application of this technique may have potential in the clinical domain, but has not been applied to date.

1.2.2.2 Linguistic features

Humans are inclined to use similar words to describe a particular concept when communicating. Linguistic features use these patterns in human natural languages including semantics and syntactics. These features can provide general contextual information that differs from medical-specific information. The most common representative linguistic feature is a part of speech (POS), which lends syntactic

information of the given sentence. UMLS Specialist Lexicon[32], MaxEnt POS tagger[42], or Stanford POS tagger[43] are medically applied automatic POS tools.

Linguistic features also have ambiguities[25, 44]. For one, linguistic tools are not always developed for clinical terminology. Therefore, medical term usage, especially acronyms and abbreviations, may be limited. Also, these tools may assume grammar consistent with standard English and may result in poor performance of these automatic linguistic tools[24, 45]. This assumption may result in errors when there are fragmented sentences in clinical notes.

Mohammad and Petersen examined the effect of lexical and syntactic features, then showed POS and other parsed features have potential to improve performances of disambiguous tasks for general English[44]. In line with the findings of Mohammad and Petersen, Coden et al. found POS was the major contributor for concept disambiguation with Mayo clinical notes[46].

1.2.2.3 Statistical features

Statistical features are based on the analysis of the given clinical corpus using statistical models. Statistical features offer domain-independent information and differ from linguistic features through the application of various statistical theories, technologies and tools. Word frequency, N-grams, Term Frequency–Inverse Document Frequency (TFIDF), window size/distance, word position/orientation, and information content are representative statistical features. TFIDF is a measure which reflects the importance of a word or term in a particular corpus[47]. Information content may have an

advantage to deal with skewed distributions of acronyms and abbreviations because it represents the mathematical quantity of information as entropy in a probability space[8]. Surrounding word collections around the target term that are disordered, are called Bag of Words (BOW). These are another powerful statistical feature[12, 25, 48] since humans tend to describe a particular concept with relatively similar words.

One disadvantage of these approaches is that it is difficult to identify various rare cases or senses because statistical features are based on fitting with statistical models. Moreover, use of more features may increase bias as the parameters of the statistical models by causing overfitting. Decisions on the cut-off point for frequency or window size are examples of additional bias. In the case of N-gram models, complex statistical models are constructed with various parameters using surrounding words within a given corpus. Lastly, the position or orientation feature of words may be used but may have problems when sentences are fragmented in clinical notes.

Ng and Lee showed the most contributing feature for their disambiguation of general English text simulation was collocation[49]. Joshi et al. also utilized collocation for general English disambiguation with high accuracy with a supervised ML algorithm[25]. In another study, Liu et al. combined BOW, orientation, distance, and collocations to solve word disambiguation with general English and biomedical documents[48]. They found that using a larger sized window provided more information to distinguish senses of abbreviations in the medical domain, which has not been found in the general English domain[48]. In another study, Pakhomov generated the training data for supervised maximum entropy models that was used to resolve six ambiguous acronyms and

abbreviations from Mayo Clinic patient notes[8].

1.2.2.4 General document features

General document features use general discourse structural information. In other words, general document features take into account cognitive flows and structures that have evolved in real clinical environments. For example, the type of the medical document, medical specialty, and structural position/orientation within clinical notes may influence the performance of these tools for clinical WSD[8, 19, 24, 50]. Among general document features, title or section information may be helpful when aggregating clinical notes.

However, clinical structural adaptation can occur because defining the structure itself is a somewhat subjective/biased issue. Furthermore, there is no guarantee that clinical notes use the same structural format not only among note types but also among EHR systems. These technologies require additional rules and technologies with domain-specific information to create the title or section database[24, 50]. Maintaining rules and databases coherently represents extra effort for management of these systems[24, 50].

Since title or section information is a distinct clinical feature, there are limited cases that it has been used. Xu et al. utilized title or section information for sense inventory creation by a clustering ML algorithm[19]. Denny et al. proposed an algorithm to categorize labeled and unlabeled titles/sections in history and physical examination notes with high accuracy[50]. This algorithm did not involve WSD per se, but the work shows the potential to use title or section information for NLP tasks.

1.2.2.5 Feature selection

NLP researchers believe that there is no single absolute best feature[12, 48], even if several representative features are critical for acronym and abbreviation disambiguation[25]. Pattern generalization is also challenging because of the already described informal nature of clinical acronyms and abbreviations. Xu et al. addressed these challenges by compounding several factors including the training sample size, the sense distribution of individual acronyms or abbreviations, and the degree of difficulty to distinguish sense meaning[51]. Generally, harmonic feature combinations without overfitting prevent skewed results and offer better performance than single feature sets[44]. With respect to general English terms, features of acronyms and abbreviations with a wider window appear to achieve better performance in the biomedical domain[48] because the medical terminology provides more specific contextual information compared to general English, where the larger window may dilute and provide extraneous information.

1.2.3 Machine learning techniques

Supervised ML algorithms are trained with annotated samples[7]. Supervised ML methods show high performance when they have enough training samples. However, it is well known that these algorithms suffer from the knowledge acquisition bottleneck. In other words, tremendous efforts, cost, and time are essential to obtain manually annotated samples by experts[19]. On the other hand, unsupervised ML techniques ("clustering") generate a model distinguishing the difference and similarity of groups of samples to one

another. Therefore, these algorithms do not need annotated senses[7]; as a result, they often have low accuracy and performance as the "learning" is not specified to the solution. Semi-supervised ML algorithms combine the advantages of supervised and unsupervised ML algorithms. In particular, these methods use minimal annotation during training and then apply unsupervised "clustering" methods to unlabeled instances. A representative example is Xu's semi-supervised approach to build sense inventories of abbreviations[19].

The degree of difficulty of WSD is inversely related to the degree of "Well-separatedness" of senses[51] to one another, which is defined as having large semantic differences between senses, and which may result in different usage in clinical notes. It has been shown that only a few dozen of well-separated samples are required during the training phase to achieve high supervised ML performance[48]. Moreover, well-separated senses tend to have low error rates for supervised algorithms[51].

Finding generalized conclusions about all clinical acronyms and abbreviations using ML algorithms may overall be difficult. To date, no absolutely superior ML algorithm has been identified to resolve ambiguous acronyms and abbreviations[25, 48]. However, in general, supervised ML methods are extensively used in WSD tasks[12, 24] because of expected high performance. Naïve Bayesian (NB) and Support Vector Machine (SVM) as supervised ML algorithms are commonly used because of their stable performance.

1.3 Specific aims, including statement of hypothesis

The overall goal of this body of work is to explore and expand automatic WSD tools for acronyms and abbreviations in the clinical domain to create new knowledge for medical NLP. To fulfill this goal, several elements have to be addressed. Figure 1.1 provides an overview of how these essential factors tie together in the development of automatic WSD tools. WSD tools need four basic inputs: preprocessed clinical documents, comprehensive clinical sense inventory(ies), optimal feature selection, and effective ML technologies. Therefore, a multi-faced approach is proposed.

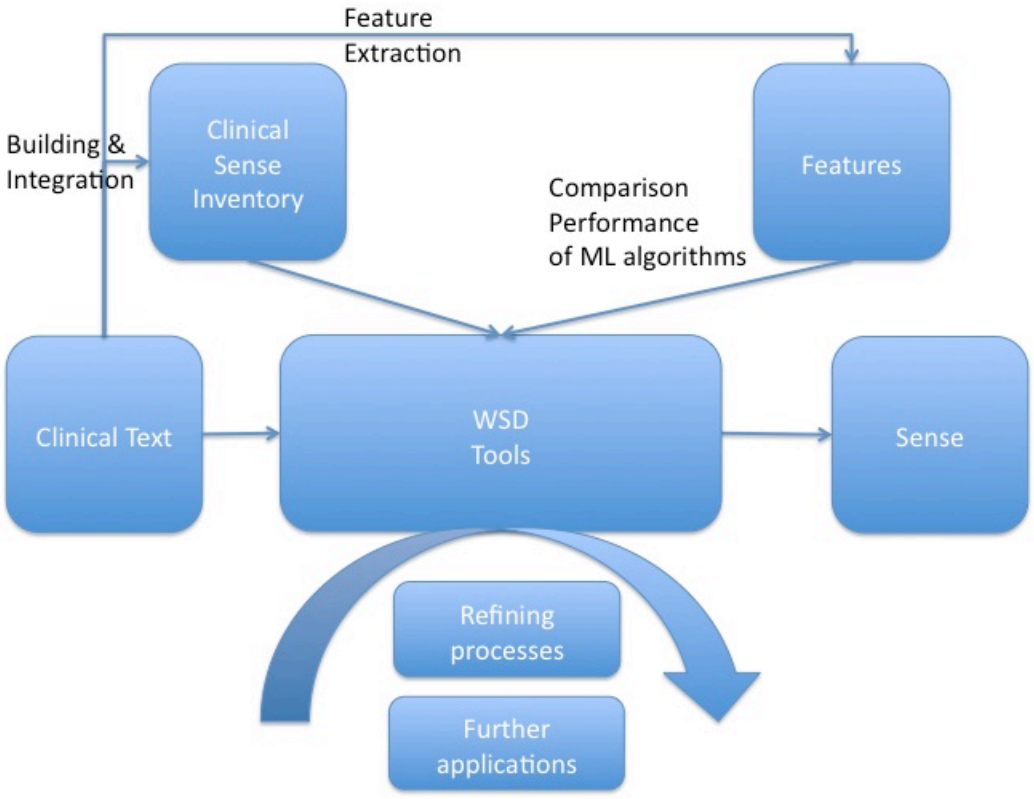

Figure 1.1 Overview of automatic WSD tools in clinical texts

First, the "Clinical Sense Inventory" (Figure 1.1) is the task of generating a clinical sense inventory. For this task, clinical acronyms and abbreviations from the corpus are harmonized with biomedical resources. Since a considerable amount of acronyms and abbreviations in the clinical fields overlap with the biomedical fields[3], taking advantage of available biomedical resources and technologies may be a significant aid in building a clinical sense inventory with limited overhead.

Optimal feature selection focuses on identifying promising features that maximize the performance of ML algorithms for WSD tools. Promising features are discussed in detail in section 1.2.2 and depicted as the "Features" (Figure 1.1). Since unnecessary information deteriorates ML algorithm performance, these algorithms often require a careful adjustment to both heterogeneous and homogeneous features. This portion of the work also requires a careful examination and comparison of ML algorithms to identify which have adequate performance, represented as the "Comparison performance of ML algorithms" (Figure 1.1). Furthermore, a minimum training sample size is needed for optimal performance.

With the "Refining processes" (Figure 1.1), it is anticipated that the quality of automatic wsd tools can be improved by examining related ares such as sense harmonization and combining utilizing biomedical resources[37] is anticipated and needed. With these refinements, WSD for clinical acronyms and abbreviations will be investigated at a deep level. However, this "Refining processes" is out of the scope of this dissertation. Another relevant issue, extension of symbol disambiguation is an area of investigation ("Further Applications", Figure 1.1) utilizing studies and technologies of

WSD for clinical acronym and abbreviation.

This dissertation presents three major topics corresponding to elements of automatic WSD tools (Figure 1.1). After the introduction and the rationale for sense disambiguation (acronym, abbreviation, and symbol) and automatic WSD tools in clinical documents in Chapter 1, these three topics are discussed in Chapter 2, 3, and 4.

A comprehensive sense inventory for acronyms and abbreviations[52] is generated and summarized in Chapter 2. This chapter accords with the "Clinical Sense Inventory" module in overview of automatic WSD tools. Senses of acronyms and abbreviations in clinical notes were manually annotated and lexically or semantically aligned with long forms of UMLS, ADAM, and Stedman's Dictionary.

In Chapter 3, this study investigates with selective 50 acronyms and abbreviations from the comprehensive sense inventory in Chapter 2, (1) optimization of the BoW window size and orientation with regard to the feature selection with various ML algorithms (2) determination of the minimum training sample size for ML algorithms[24]. These investigations are corresponding to the "Features" and "Comparison Performance of ML algorithms" module in Figure 1.1.

Based on the studies in Chapter 2 and 3, WSD technologies are then extended to automatic non-alphanumeric symbol resolution[53] in clinical notes in Chapter 4. Similar to Chapter 2, annotated senses of four common symbols ('+', '–', '/', and '#') in clinical notes are compared with senses from linguistic literatures, medical literature, and Stedman's dictionary. Using analogous principles to Chapter 3, extracted features for symbols are utilized with various classifiers to perform symbol sense disambiguation as

well. Studies in Chapter 4 conform to the "Further applications" box in Figure 1.1. Lastly, results of these three studies in Chapter 2, 3, and 4 and suggested future directions are presented in Chapter 5.

Chapter 2 A COMPREHENSIVE SENSE INVENTORY FOR CLINICAL ABBREVIATIONS AND ACRONYMS USING BIOMEDICAL, BIOMEDICAL LITERATURE, AND MEDICAL DICTIONARY RESOURCES[1]

Sungrim Moon[1], Serguei Pakhomov[1,2], Nathan Liu[3], James O. Ryan[1], Genevieve B. Melton[1,3]

[1]Institute for Health Informatics; [2]College of Pharmacy; [3]Department of Surgery

University of Minnesota, Minneapolis, MN, USA

Objectives: To create a comprehensive sense inventory of abbreviations and acronyms from clinical texts.

Design: The most frequently occurring abbreviations and acronyms from 604,944 dictated clinical notes were used to create a clinical sense inventory. Senses of each abbreviation and acronym were manually annotated from 500 random instances and lexically matched with long forms within the Unified Medical Language System (UMLS Version 2011AB), Another Database of Abbreviations in Medline (ADAM), and *Stedman's Dictionary, Medical Abbreviations, Acronyms & Symbols, 4th edition* (*Stedman's*). Redundant long forms were merged after they were lexically normalized

[1] This work was supported by the University of Minnesota Institute for Health Informatics Research Support Grant (SM, GM, SP), the American Surgical Association Foundation Grant (GM), by the National Library of Medicine (#R01 LM009623-01) (SP). We would like to thank Fairview Health Services for support of this research. The authors also thank Andrea Abbott for manual annotations, and Bridget McInnes for insightful comments.

using Lexical Variant Generation (LVG).

Results: The clinical sense inventory was found to have skewed sense distributions, practice-specific senses, and incorrect uses. Of 440 abbreviations and acronyms analyzed in this study, long forms for 949 were identified in clinical notes. This set was mapped to 17,359, 5,233, and 4,879 long forms in UMLS, ADAM, and *Stedman's* respectively. After merging long forms, only 2.3% matched across all medical resources. The UMLS, ADAM, and *Stedman's* covered 5.7%, 8.4%, and 11% of the merged clinical long forms respectively. The sense inventory of clinical abbreviations and acronyms and de-identified datasets generated from this study are available for public use at http://purl.umn.edu/137703 (website).

Conclusion: Clinical sense inventories of abbreviations and acronyms created using biomedical, biomedical literature, and medical dictionary resources demonstrate challenges with term coverage and resource integration. Further work is needed to help with standardizing acronyms and abbreviations in clinical care and biomedicine to facilitate automated processes such as text-mining and information extraction.

2.1 Introduction

Abbreviations and acronyms in biomedical and clinical documents are pervasive, and their use is expanding rapidly[1-3, 8, 51]. With the accelerated adoption of electronic health record (EHR) systems and proliferation of clinical texts, there is an increasing need to deal with abbreviations and acronyms and to utilize electronic clinical documents for automated processes. In addition to electronic clinical notes that are traditionally

created by dictation and transcription, many clinical notes are now created at the point of care where clinicians type, dictate using voice recognition software, enter notes using a semi-structured or templated document entry system, or use a hybrid of several of these approaches. This often results in the use of shortened word forms that often have multiple meanings and may present a challenge for subsequent automated information extraction from notes and also may potentially result in patient safety issues[5, 6, 54].

Sense inventories of abbreviations and acronyms are important and considered an essential component for automated natural language processing (NLP) systems. Abbreviation and acronym sense resolution, a special case of word sense disambiguation (WSD)[7, 9, 27], is most effectively achieved based on the presence of a consistent and complete sense inventory. Compiling sense inventories is a challenge, however, since they are labor intensive and work to date in the clinical domain is somewhat limited resulting in limited availability of clinical sense inventories.

Although abbreviation and acronym sense inventories have been studied extensively for biomedical texts specifically within the biomedical literature, relatively little research has been devoted to the creation of a sense inventory of abbreviations and acronyms within clinical notes[19, 25]. With biomedical literature[14-17, 21, 22], typically the first instance of a short form for the abbreviation or acronym occurs with the long form as a parenthetical expression or vice-versa (e.g., "mucosal ulcerative colitis (MUC)")[13]. In contrast, clinical notes are informal in nature and the association of long form and short form in clinical text is rarely observed[19, 51]. Moreover, the development of any abbreviation and acronym sense inventory from clinical texts is hindered by issues of

patient confidentiality and privacy that make sharing and using clinical notes for research purposes difficult[2, 12]. Not surprisingly, there are currently only small clinical sense inventory datasets of abbreviations and acronyms available (e.g., datasets by Xu et al.[19] or the Mayo Clinic[25]).

The goal of this work is to create and release for public use a comprehensive clinical sense inventory of clinical acronyms and abbreviations, harmonized with a medical dictionary *Stedman's Medical Abbreviations, Acronyms & Symbols, 4th edition* (*Stedman's*)[55]; the Unified Medical Language System (UMLS)[56]; and an acronym and abbreviation sense inventory from biomedical literature, Another Database of Abbreviations in Medline (ADAM)[21] From this work, we sought to understand different usages of clinical abbreviations and acronyms and the relative coverage and degree of overlap across these resources.

2.2 Background

2.2.1 UMLS

The UMLS is distributed through the National Library of Medicine as a set of medical terminology resources organized by concepts. In addition to providing a resource for identification of medical terms, the UMLS provides ontological information for concepts based upon an "is-a" hierarchy, includes lexical variants for concepts, has source terminology mappings, and other types of relationships between concepts (e.g., "treats")[12, 27, 28]. While the UMLS is a natural resource for mapping senses of clinical abbreviations and acronyms, the UMLS has previously been shown to have limited

coverage of acronyms and abbreviations[39] although some work has shown improved coverage for a subset of acronyms. For example, Xu et al.[3] in 2007 found that the UMLS only covered approximately 35% of the abbreviations and acronyms that the authors examined in the clinical domain. Similarly, Liu et al.[4] reported coverage of 66% of examined abbreviations and acronyms with less than 6 characters in the clinical domain by the UMLS.

There are a number of relational files and tools available to access and utilize the UMLS. For example, the National Library of Medicine provides the Specialist Lexicon[32] (including the LRABR file) and a part of the Specialist Lexicon tool, Lexical Variant Generation (LVG)[57], which allows for term normalization and stemming in the distribution of MetaMap[33]. Moreover, MetaMap, which was used in this study, is a software application developed to map text to corresponding biomedical concept(s) indexed with the UMLS concept unique identifier (CUI) and its associated UMLS semantic type (the UMLS semantic type of each concept).

2.2.2 ADAM

A number of rule-based and statistically generated sense inventories have been created using the assumption that the short form and the long form of an abbreviation or acronym are collocated when first introduced in biomedical literature documents (e.g., SaRAD[14], ARCH[15], and ALICE[16]). Among them, ADAM is a representative acronym and abbreviation biomedical sense inventory resource generated from titles and abstracts via 2006 MEDLINE[21]. ADAM contains 59,403 pairs of short and long forms as a database for B-terms projected after filtering out insignificantly connected pairs based on

length ratio rules and empiric cut-off values. ADAM also provides the term frequency of different terms along with other statistical information to illustrate usage of each abbreviation or acronym within the biomedical literature. ADAM does, however, contain a significant level of redundancy between different long form expressions owing to the lack of work to perform either syntactic or semantic normalization between different expressions.

2.2.3 Medical dictionaries

Medical dictionaries such as *Stedman's* and *Dorland's* are currently not available as part of the UMLS and thus tend to be underutilized in the development of biomedical and clinical NLP work. These dictionaries may, however, provide an important adjunctive resource for clinical sense inventories because medical dictionaries are used commonly within the clinical domain and have a large amount of information about biomedical and clinical terms represented in texts. The definitions of terms in these resources can also be potentially used to constrain semantic information for related tasks such as word sense disambiguation[58]. On the other hand, potential issues with medical dictionaries include copyright restrictions, the comparative slowness of these resources to adopt new clinical terms, and the hybrid nature of these resources, which contain both clinical as well as basic science terms.

2.3 Method

Clinical documents from four hospitals in the University of Minnesota-affiliated Fairview Health Services, including the University of Minnesota Medical Center and other Fairview metropolitan hospitals in the Twin Cities, from 2004 to 2008 in our clinical document data repository were used for this study. The corpus contains primarily verbally dictated and transcribed notes stored in electronic format. These 604,944 clinical notes include admission notes, operative reports, consultation notes, and discharge summaries.

2.3.1 Identification of significant abbreviations and acronyms

To select meaningful and common abbreviations and acronyms, a set of heuristic rules were applied. Potential abbreviations and acronyms were identified when the word token consisted of capital letters or numbers, with or without symbols (period, comma, colon, or semicolon) using regular expressions. Combinations of symbols in front or in back of the targeted word token were accepted as a potential abbreviation or acronym. If the token of interest was part of document formatting (e.g., header, footer, or transcription formatting), it was excluded. Heuristic rules were applied to clinical notes to detect the section information for the abbreviation or acronym. Only candidate abbreviations or acronyms with a frequency of over 500 in the corpus were included resulting in 440 abbreviations and acronyms. The surrounding text for each of the 500 instances was also extracted and included in the inventory. The instance consisted of 12 previous-word tokens and 12 post-word tokens centering the targeted abbreviation and

acronym. A set of twelve word tokens was selected based upon previous work in general English showing that this is more than sufficient for manual annotation[10].

2.3.2 Identification of possible long forms from various medical areas

All 220,000 instances for the 440 abbreviations and acronyms were given to two clinical experts for manual annotation of their clinical sense. Annotated long forms were then standardized with long forms of *Stedman's Medical Abbreviations, Acronyms & Symbols, 4th edition* (*Stedman's*). We choose *Stedman's* among Medical dictionaries because this was available electronically and had a resource specific for abbreviations and acronyms. At this stage, formatting errors were eliminated and replaced by additional samples focusing the clinical sense inventory upon the overall sense distributions of our corpus. For example, "1. Atrial fibrillation. 2. C3. omfort cares…", 'C3' is not a valid abbreviation or acronym but rather a formatting mistake. The inter-rater reliability of the annotated senses was reported with percentage agreement and with the Kappa statistic with a third clinical expert who examined 11,000 random samples (25 per abbreviation or acronym – 5% of the total).

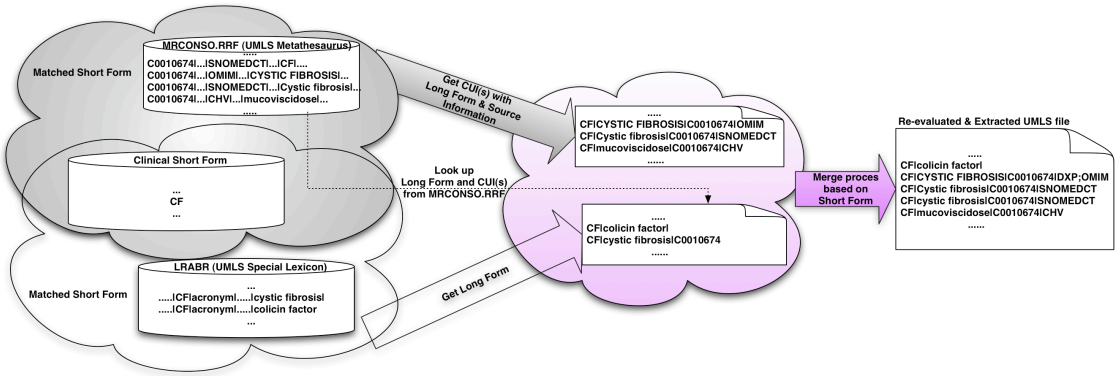

Figure 2.1 Collect long forms from the UMLS

Figure 2.1 provides an overview of how potential long forms in the UMLS were obtained for each of the acronyms and abbreviations. As a first step, each short form of a given clinical abbreviation or acronym was mapped using the Metathesaurus file MRCONSO.RRF (UMLS 2011AB) to determine the corresponding long form(s), CUI(s) and English term type(s) (shaded box in figure and arrow). Second, the given clinical short forms were mapped using the LRABR file to extract pairs of short forms and long forms mapped in the UMLS. These long forms from the LRABR file (the UMLS Specialist Lexicon) were re-mapped to MRCONSO.RRF to get CUI(s) and English term type(s) (dotted line). Third, all identified long forms from the first and second steps were merged based on short forms ("Merging process based on Short Form"). Fourth, collected CUIs and long forms were remapped (dotted line) to MRCONSO.RRF one more time to detect any missing variants of the long forms/information in the UMLS. The result of this process is represented as "Re-evaluated & Extracted UMLS file" in Figure 2.1.

Short forms of abbreviations and acronyms in the clinical domain were directly

mapped to short forms of ADAM since ADAM has paired representations of short forms and long forms of abbreviations and acronyms. Additionally, we included the coverage and usage frequency of individual long forms from ADAM so as to include information about the relative usages within the biomedical literature.

Finally, for each short form, all long forms associated with a targeted clinical abbreviation or acronyms were extracted from *Stedman's*. All bracketed expressions in the dictionary were reviewed to select all possible inflected forms. For example, "TEE" had an original representation as "transesophageal echocardiograph(y) (echocardiogram)" in *Stedman's*. For this, "echocardiogram" and "echocardiography" were kept because they have similar meanings to "echocardiograph". As a result, we had three expressions for "TEE": "transesophageal echocardiogram", "transesophageal echocardiograph" and "transesophageal echocardiography".

2.3.3 Normalization process and analysis of the sense inventory

The initial sense inventory for the source clinical abbreviations and acronyms was systematically compared to each of the resources (UMLS, ADAM, *Stedman's*) to identify similarities and differences. Figure 2.2 provides an overview of the mapping processes for all acquired long forms from various medical resources. A two-step process was used to merge long forms by applying a lexical step followed by a semantic merging step.

Before the two-step process, all previously obtained long forms were used as inputs into MetaMap. Concept Unique Identifiers (CUIs) produced by MetaMap as final mappings were included only if they had a score of 1,000 (highest score/confidence) to

ensure exact mapping of given long forms. MetaMap term processing option (-z) was used to obtain exact matching when MetaMap processed long forms. The "-z" term processing option makes it so that MetaMap deals with the individual chunk of strings as a single phrase/unit (rather than a sentence or a full text). Therefore, MetaMap processes inputs without applying the split process, which helps to obtain suitable mappings for the vocabulary terms. Each identified long form has a relevant set of CUI(s) (from MRCONSO and MetaMap) that was included in the inventory.

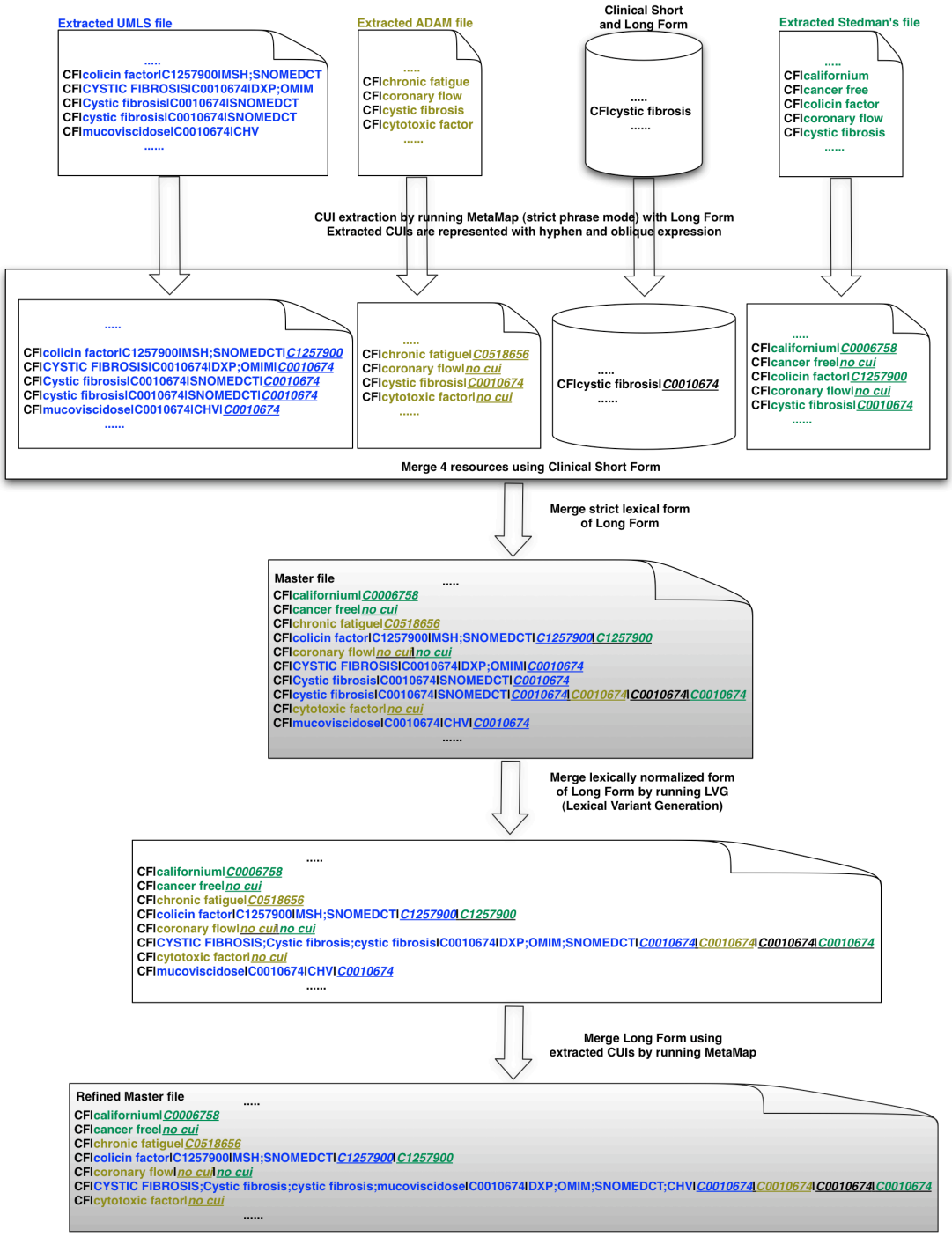

Figure 2.2 Merging process of long forms
Extracted UMLS file = result from Figure 1, UMLS = The Unified Medical Language System, ADAM = Another Database of Abbreviations in Medline, Stedman's = Stedman's Medical Abbreviations, Acronyms & Symbols, CUI = Concept Unique Identifier

Lexical merging of long forms was first performed to find exact matches of lexical forms of each acronym's long form in various medical resources. Only long forms with the same lexical representations to each other were used to create the "Master file" as shown in Figure 2.2. Following this, LVG normalization with individual long forms was used to remove simple variations of lexical representations. Examples of these simple variations of lexical representations include plural expressions, word orders differences, existence of stop words (e.g., "and", "the", "of"), and variation in punctuation and other symbols. Long forms with exactly the same normalization through LVG were then merged as one concept.

Following lexical matching, semantic mapping was performed based on CUIs between long forms to enhance the quality of sense inventory. Only perfect mappings based on CUIs from UMLS were taken into consideration. In other words, if any set of CUIs for given long form have an overlap of 100% to the set of CUIs for another long form, the two long forms were regarded as the same concept/meaning but had the different lexical representations. These semantically equivalent long forms were mapped into as one representation in our "Refined Master file" as shown in Figure 2.2.

2.4 Result

2.4.1 Characteristics of clinical sense inventory

Within the overall clinical corpus of 604, 944 notes, 440 common abbreviations and acronyms with 949 long forms were found occurring with a frequency of 500 of more instances in the corpus. For inter-rater reliability, the percent agreement was on average

99% and the Kappa statistic was on average 0.97 between annotators. Among acronyms and abbreviations, GTT (80%, 0.25), SI (84%, 0.30), GT (84%, 0.30), US (76%, 0.35), NP (88%, 0.36), INH (88%, 0.47), ES (92%, 0.48), PCA (92%, 0.48), AP (96%, 0.49), and DP (96%, 0.49) had fair to high percent agreement and low Kappa statistic respectively.

Table 2.1 Kappa statistic in clinical corpus

Range of value of Kappa	Number of abbreviations and acronyms
0.90 – 1.00	398
0.80 – 0.90	16
0.70 – 0.80	10
0.60 – 0.70	6
Less than 0.60	10
Total	440

Table 2.2 Sense distributions in clinical corpus

Ratio of majority sense	Number of abbreviations and acronyms
99 – 100%	323
95 – 99%	42
90 – 95%	14
80 – 90%	21
70 – 80%	11
60 – 70%	8
50 – 60%	14
Less than 50%	7
Total	440

The great majority of abbreviations and acronyms in the clinical sense inventory had skewed distributions for meanings. Overall, 276 of 440 (62.7%) of abbreviations and acronyms had only a single sense (long form). This majority sense prevalence was significantly different in comparison to the distributions seen in the biomedical literature. Table 2.2 shows the frequency distribution of the clinical senses sorted according to the

baseline majority sense rate. Of all cases, 83% of had one dominant majority sense using a conservative ratio of >95% as the definition of a dominant majority. The clinical sense inventory contained several institution-specific terms with senses that were not generalizable to the greater clinical domain. For example, in Table 2.3, the acronym "FUTS" is a short form for "Fairview University Transitional Services." Another similar example, "FSH" in the dataset was often used (46%) to represent "Fairview Southdale Hospital."

Table 2.3 Sense of FUTS and FSH

Abbreviation	Sense	Number of instance	Coverage
FUTS	Fairview University Transitional Services	500	1.00
FSH	follicle-stimulating hormone	265	0.53
	Fairview Southdale Hospital	231	0.46
	fascioscapulohumeral muscular dystrophy	4	0.01

Overall, 335 cases of misuse of acronyms were observed in corpus used to create the clinical sense inventory. For example, the text in one instance stated: "…PAC pump for anesthesia…" which should have been "PCA (patient-controlled analgesia)" rather than "PAC". In another example: "The patient is on Biaxin for mycobacterium AVM intracellular infection". Here, "AVM" was misunderstood and should have been the word "avium" and should have been a word, the mistake occurring in the context of dictation and transcription. Most frequently in our dataset, we observed mistaken use of "BMP" which should have been "BNP" 36 times, "BNP" which should have been "BMP" 18 times, "DT" which should have been "DP" 23 times, and PM which should have been "PMR" 74 times.

An additional 306 errors were observed. An example of a mistake with unclear

meaning includes the following: “His factor 2 SA was 14 on admission and factor 12 SA was 62.” We represented these unsure cases as “unsure sense” in our clinical document repository. Sometimes, the detected abbreviation or acronym was a part of word phrase that together had a particular meaning. For example, “Mucolytics and EC PAP device.” “EC PAP” should be corrected as “ EZ PAP” but “EZ” itself has no meaning without “PAP”.

2.4.2 Comparison among different resources

Figure 2.3 represents the coverage among resources. Looking only at those long forms with an exact match of lexical forms, among 24,853 total senses (long forms) of 440 abbreviations and acronyms, 224 total were matched exactly between all resources. For example, the abbreviation ABG had a single sense “arterial blood gas” with the CUI “C0150411”. All sources (UMLS, ADAM, and *Stedman’s*) had the long form “arterial blood gas”. Some long forms represented several preferred CUIs, like AVM had the sense “arteriovenous malformation” with two associated CUIs: “C0003857” and “C0334533”. Overall, these exact and completely matched long forms for all medical resources represented only 0.9% of long forms in the dataset (224 of 24,853 long forms).

The low rate of matching long forms across all resources was improved after the processing and merging of long forms. 24,853 initial total long forms were merged into 17,096 long forms after performing LVG normalization (Figure 2.2). At this stage, exact and complete match of long forms for all resources increased to 1.7% (296 of 17,096 long forms). After we applied semantic matching for equivalent CUIs, the exact match

rate increased to 2.3% (302 of 13,386 long forms).

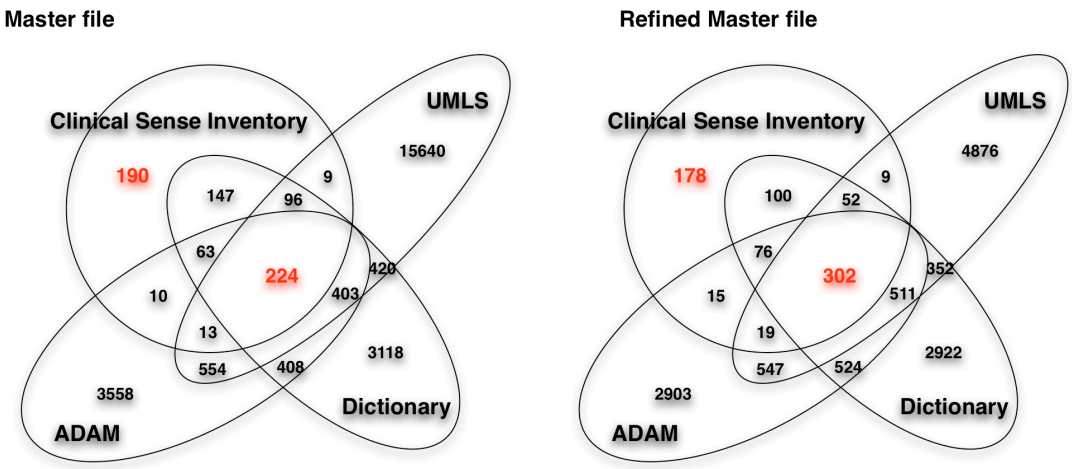

Figure 2.3 The coverage among resources
Master file and Refined Master file = result from Figure 2, UMLS = The Unified Medical Language System, ADAM = Another Database of Abbreviations in Medline, Stedman's = Stedman's Medical Abbreviations, Acronyms & Symbols

After the three-phrase merge process, clinical long forms covered 50.9% (382 of 751 long forms) of the UMLS, 54.9% (412 of 751 long forms) of ADAM, and 70.6% (530 of 751 long forms) in *Stedman's*. When looking at the coverage relative to the clinical sense inventory, the coverage of UMLS, ADAM and *Stedman's* was 5.7% (382 of 6,668), 8.4% (412 of 4,897), and 11% (530 of 4,839) respectively of long forms in the clinical sense inventory.

We also observed that the use of abbreviations was different between the clinical and biomedical domain, specifically when comparing the clinical sense inventory with ADAM. For example, ODT is used (100%) for "orally disintegrating tablet" in our clinical sense inventory but in the biomedical literature, ODT is used (100%) for

"Oculodynamic Test". Similarly, FEN means (100%) for "fluids, electrolytes, nutrition" for in clinical domain, but it is mainly used (68.1%) for "fenfluramine"(C0015827) in biomedical literature. We found different usage by domain (100% dominantly used in clinical sense inventory but less than 50% in ADAM) with 33 acronyms and abbreviations.

We observed that some clinical senses have did not correspond to long forms within any of the resources. Among 949 long forms in the clinical sense inventory, 190 had no coverage in any of the three resources using exact matches of lexical forms. This was reduced through LVG normalization and semantic matching, which reduced the number of unmatched long forms to 178. Table 2.4 gives some examples of long forms among the four resources.

Table 2.4 Sense comparisons between the clinical sense inventory and other resources

Short Form	Long Form	MetaMap CUI	CSI	Ratio in CSI	UMLS CUI	UMLS SOURCE	ADAM	Ratio in ADAM	Stedman's
AVR	Lead AVR;lead avr;aVR	C0449217			C0449217	CHV;LNC;SNOMEDCT			
	aVR (body structure)				C0449217	SNOMEDCT			
	accelerated ventricular rhythm								1
	acute vascular rejection						1	0.0634	
	antiviral regulator								1
	aortic valve regurgitation	C0003504	1	0.01					
	aortic valve replacement	C0003506	1	0.762	C0003506	CHV;COSTAR;NCI;SNOMEDCT	1	0.8687	1
	aortic valve resistance		1	0.008					
	arteriole-to-venule ratio						1	0.0133	
	ascending vasa recta	C2952018			C2952018	FMA	1	0.0398	
	augmented voltage right arm		1	0.206					1
	pathogen avirulence						1	0.0147	
BKA	Bka;CGI-35;FCF1;FCF1 gene	C1426785			C1426785	HUGO;MTH			
	FCF1 small subunit (SSU) processome component homolog (S. cerevisiae)				C1426785	HUGO			
	below-knee amputation	C0002692	1	1	C0002692	NCI	1	0.5714	1
	bongkrekic acid	C0005982			C0005982	MSH;NDFRT	1	0.4286	
IOL	IOL;iol;Primary Intraocular Lymphoma	C0281658			C0281658	CHV;NCI;PDQ			
	Intraocular Lymphoma;Intraocular lymphoma;intraocular lymphoma;lymphoma, intraocular;intraocular lymphoma (IOL)	C0281658; C1706527				CHV;MTH;NCI;PDQ			
	induction of labor	C0259787							1
	interosseous ligament	C0447892					1	0.0968	
	intraocular lens;intraocular lenses	C0023311; C0023319; C0336564	1	1	C0023319; C0336564; C1706007	CHV;MSH;NCI;SNOMEDCT	1	0.7849	1
	intraocular lens implantation	C0023311					1	0.1183	

UMLS = The Unified Medical Language System, ADAM = Another Database of Abbreviations in Medline, CSI = Clinical Sense Inventory, Stedman's = Stedman's Medical Abbreviations, Acronyms & Symbols, CUI = Concept Unique Identifier, UMLS SOURCE = Source information in the UMLS, MetaMap CUI = CUI produced by running MetaMap

2.5 Discussion

This study provides and evaluates a comprehensive sense inventory for clinical abbreviations and acronyms and compares and contrasts the long forms and short forms amongst three resources: UMLS, ADAM and *Stedman's*. The clinical sense inventory had overall highly skewed sense distributions, some local or practice-specific senses, and a number of erroneous instances. Our analysis of the 440 most common abbreviations and acronyms from clinical notes demonstrated that many long forms were not perfectly matched even after conducting lexical mappings and semantic comparisons. Despite some of the challenges and limitations encountered in the process of creating the sense inventory, we believe that the resultant resource from this study currently represents the largest and most comprehensive sense inventory of clinical acronyms and abbreviations. This resource is publically available to support the research of the greater NLP and biomedical health informatics community.

We observed that vocabulary resources used in this study had uneven granularity of sense distributions as compared to each other. This created challenges in the normalization process of the inventory's long forms. For example, ADAM and the UMLS distinguished "total knee arthroplasty" and "total knee arthroscopy". In contrast, *Stedman's* collapses these two concepts in a single sense: "total knee arthroplasty (arthroscopy)". Because this combined sense is not suitable for obtaining CUIs with MetaMap and has two semantic meanings, this was separated into two expressions for our study.

Another challenge encountered with the sense inventory was that of ambiguous abbreviations or acronyms within the text. For example, "Imdur SA 60 mg p.o. q.d." "SA" can be either "slow acting" or "sustained action", which has a similar sense but different long form expansions. The occurrence of two meaningful senses repeatedly occurring was prominent in a few abbreviations/acronyms. These ambiguous senses were observed 373 times with "SA" ("slow acting" or "sustained action"), 121 times with "OP"("oblique presentation" or "occiput posterior") and 105 times with "MP"("metatarsophalangeal" or "metacarpophalangeal") in the 500 samples of those particular abbreviations/acronyms. We also observed some ambiguity associated with senses related to levels. For example, abbreviation "C3" has one representative sense "cervical (level) 3". Here "level" can be interpreted one of several meanings such as "nerve", "dermatome", "vertebrae" or "disc" depend on surrounding words.

Another issue with term normalization across resources was the degree of redundancy of long form terms, particularly the significant degree of redundancy in ADAM amongst its different long forms, where all distinct lexical forms remained separated. Additional steps are required to further reduce the redundancy of long form senses prior to mapping to ADAM long forms to other resources. While some work has been done to merge synonymous variants of the long forms[37], our sense inventory only utilized strict and exact matching processes.

The assumption used in biomedical literature and general English is generally that there is only one sense per discourse per abbreviation/acronym. This assumption stems from NLP work in general English word sense disambiguation[59]. We found this to be an

invalid assumption for clinical documents. In some instances, even the assumption of one sense per sentence does not hold in clinical discourse making the problem of word sense ambiguity resolution more challenging in this domain. We found several examples where two senses for an acronym/abbreviation were observed within a single sentence such as: "Postop MRI recently showed increase T2 signal from C2 through T2 level." Here, the first "T2" means "T2 (MRI phase)" but the second "T2" means "thoracic vertebra 2". We did find, however, that most instances of "T2 (MRI phase)" appeared in the section, "PROCEDURE", and the sense "thoracic vertebra 2" appeared mostly in the section "HISTORY OF PRESENT ILLNESS", indicating that the section may be helpful for determining the sense of an abbreviation/acronym in a clinical discourse. The section information will not be helpful in all cases, however.

One observation that has been made previously[3, 4] and confirmed by our study is that the UMLS is limited as a resource for mapping short forms with long forms. The LRABR file in the UMLS contains overall 57,704 pairs of short and long forms. Of the 949 long forms, 190 in the clinical sense inventory were missing in the UMLS. This fact demonstrates challenges. With *Stedman's* and ADAM, there was less coverage of long forms although some other areas of coverage not afford by the UMLS, pointing to the complementary nature of these resources.

Our study has several additional limitations. After performing exact lexical matching, the techniques used for normalization of senses were dependent upon the automated tools we used (i.e., MetaMap and LVG), which may introduce additional errors in the normalization process. Because our sense inventory was built based upon

only 500 random samples that were extracted and manually annotated, these samples may not be completely representative of the entire corpus. It is also possible that these samples exclude additional minority senses. In future work, we plan to utilize semi-automated methods as previously described[19] to enrich our sense inventory, concentrating our effort on abbreviations/acronyms without a single dominant sense. Nevertheless, this study and the its associated resultant sense inventory represents a significant contribution and resource for others to use in the clinical NLP domain. The de-identified dataset of acronyms and abbreviations (those with dominant sense <95%) and sense inventories are publically available at http://purl.umn.edu/137703 (website).

2.6 Conclusion

Although abbreviations and acronyms in clinical text are used widely in clinical documentation, relatively little work has focused upon building a comprehensive clinical sense inventory for abbreviations and acronyms for the purposes of NLP research and dissemination to the wider scientific community. We created a clinical sense inventory with 440 common abbreviations and acronyms and compared the senses with the UMLS, ADAM, and *Stedman's*. From this, we were able to examine the information within and perform a gap analysis of these clinical acronyms and abbreviations among diverse resources.

Chapter 3 AUTOMATED DISAMBIGUATION OF ACRONYMS AND ABBREVIATIONS IN CLINICAL TEXTS: WINDOW AND TRAINING SIZE CONSIDERATIONS[2]

Sungrim Moon, MS[1], Serguei Pakhomov, PhD[1,2], Genevieve B. Melton, MD, MA[1,3]

[1]Institute for Health Informatics, [2]College of Pharmacy, [3]Department of Surgery

University of Minnesota, Minneapolis, MN

Acronyms and abbreviations within electronic clinical texts are widespread and often associated with multiple senses. Automated acronym sense disambiguation (WSD), a task of assigning the context-appropriate sense to ambiguous clinical acronyms and abbreviations, represents an active problem for medical natural language processing (NLP) systems. In this paper, fifty clinical acronyms and abbreviations with 500 samples each were studied using supervised machine-learning techniques (Support Vector Machines (SVM), Naïve Bayes (NB), and Decision Trees (DT)) to optimize the window size and orientation and determine the minimum training sample size needed for optimal performance. Our analysis of window size and orientation showed best performance using a larger left-sided and smaller right-sided window. To achieve an accuracy of over 90%, the minimum required training sample size was approximately 125 samples for SVM classifiers with inverted cross-validation. These findings support future work in

[2] The authors would like to thank Fairview Health Services for ongoing support of this research. This work was supported by the University of Minnesota Institute for Health Informatics Research Support Grant (SM, GM, SP), the American Surgical Association Foundation Grant (GM) and the National Library of Medicine (#R01 LM009623-01) (SP).

clinical acronym and abbreviation WSD and require validation with other clinical texts.

3.1 Introduction

Acronyms and abbreviations within clinical texts are widespread, and their use continues to increase[1, 3, 12]. Several reasons for this ongoing growth include adoption of electronic health record (EHR) systems with increased volume of electronic clinical notes accompanied by the wide usage of acronyms and abbreviations[3], the time-constrained nature of clinical medicine encouraging the use of shortened word forms, and a longstanding tradition of commonly using acronyms and abbreviations in clinical documentation[1]. The process of understanding the precise meaning of a given acronym or abbreviation in texts is one of several key functions of automated medical natural language processing (NLP) systems[9] and is a special case of word sense disambiguation (WSD)[2]. Automatic meaning discrimination by a machine is a complex task that is critical to accessing information encoded in clinical task[7, 10]. Improved acronym and abbreviation WSD methods can therefore enhance automated utilization of clinical texts to support diverse applications that rely on NLP.

Acronyms and abbreviations each have a short form (the acronym or abbreviation) and a long form (the expansion of the acronym or abbreviation). In clinical documents, the expanded long form is rarely proximal to the short form of the acronym or abbreviation[3, 19] because clinical texts rarely conform to the formalism of enclosing the long form in parentheses after the first mention of the abbreviation, as is customary in scientific literature[13]. This lack of the formalism is one of the significant barriers

associated with using clinical texts for NLP research, which has resulted in limited data resources for research. Because of this informality and the shortage of the available resources/research, while researchers have explored the use of supervised machine learning (ML) approaches for acronym and abbreviation WSD[2, 12, 25], some of the related issues with optimal window size and orientation and with training sample size minimization to reduce the associated cost and time to manually annotate training corpora remain open[12, 25].

In this paper, we have three objectives: (1) to understand and validate the relative value of different features to automatically disambiguate senses of 50 clinical acronyms and abbreviations; (2) to determine the optimal window size and orientation for obtaining features for acronym and abbreviation sense disambiguation; and (3) to estimate minimum sufficient training sample size for good performance in the inverted cross-validation settings using supervised learning approaches.

3.2 Background

3.2.1 Broad classes of features for WSD

Types of predictive features from clinical notes can be grouped into domain knowledge-based, linguistic, statistical, and general document features. These features utilize techniques developed in the biomedical NLP and computational linguistics domains. Optimal feature selection for WSD therefore requires a comprehensive understanding of the strengths and weaknesses of each feature type to maximize valuable information used for feature sets as input into ML algorithms.

Because clinical notes are based upon medical knowledge, biomedical and clinical domain resources can serve as the knowledge base to enhance clinical WSD algorithms. In particular, the Unified Medical Language System (UMLS)[23] and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT)[26] are terminology resources in the biomedical and clinical domains respectively. These resources are used by the medical NLP community not only because they provide knowledge sources for identification of medical terms, but also because they offer semantic information and ontological relationships[9, 12] that may be used to compute semantic similarity measures between concepts that can subsequently serve as features for ML[58]. On the other hand, while medical terminologies have face-validity of concept coverage, the curation and quality of these resources are variable for different subject domains and must be considered in any error analysis involving the use of these resources[31]. Automatic tools for the UMLS have been used with success in biomedical WSD research. For example, MetaMap automatically maps terms in texts to biomedical concepts of the UMLS[33]. McInnes et al.[39] showed that Concept Unique Identifiers (CUIs) generated by MetaMap to biomedical concepts of the UMLS to be good features for general WSD using supervised ML algorithms in the biomedical domain. Leroy and Rindflesch[28, 60] examined semantic types and groups with MetaMap. These ontology features produced high variability in accuracy of supervised ML algorithms, because these features rely on complicated hierarchical semantic knowledge representation and have low granularity[34].

Linguistic features are based upon patterns of human natural languages and are applicable to clinical notes that result from human communication in the clinical domain.

These features represent characteristics of language in general and reflect structural properties of a particular language that are independent of the medical domain. Automatic tools, like the Stanford part-of-speech (POS) tagger[43] or MedPOST[61], have been trained on general English and biomedical discourse, respectively. However, these tools may not always perform well with clinical texts due to frequent deviations from the standard English sentence structure[45]. The most common linguistic feature set used in WSD is POS information, which indicates the syntactic category of a given word as it is used within a sentence. Mohammad and Peterson[44] utilized lexical and syntactic features to improve performance of supervised classifiers for general English WSD.

Statistical features utilize distribution and co-occurrence of features of a given corpus. Because humans often describe ideas with similar words, these features are powerful and supported through well-established statistical theories, technologies, and tools. However, one of the weaknesses of these approaches is the difficulty of detecting rare cases or minor senses. In contrast, parameters of statistical models can increase bias through overfitting. Bag-of-words (BoW) is the simplest example of using the frequency of lexical items surrounding the ambiguous word as a predictive statistical feature[62]. Despite its apparent simplicity and a number of limitations, BoW approach has been demonstrated in previous studies to provide high quality information for many WSD tasks[12, 25]. Joshi et al.[25] explored BoW and term frequency applying supervised approaches to improve accuracy of ML algorithms. Liu et al.[48] investigated diverse feature sets including BoW with 15 biomedical abbreviations with supervised ML algorithms. In this later study, the authors show that BoW or BoW with a few word-based

features (corresponding orientation within a three word windows with three nearest two-word collocations) produce the best performance for abbreviation disambiguation.

Finally, general document features include information related to the global discourse structure (e.g., document title or section headings). Document characteristics may be indicative of a type of the medical document or clinical sub-specialty and may help narrow down a particular rule set for a particular NLP task[50]. Discourse information therefore incorporates idiosyncrasies of clinical documentation into predictive features for WSD. For instance, Xu et al.[19] used in part section information to build 12 sense inventories from a repository of admission notes through semi-supervised ML methods. One limitation of this class of features is that clinical notes do not always use the same structural format for the same note type, even within the same hospital system or same EHR system. This set of features may also require significant domain knowledge and development of specific rules based upon context, also resulting in a large overhead and lower scalability[50].

3.2.2 Feature selection considerations

Even though researchers have used diverse approaches for WSD, limited studies in the clinical domain make optimal feature sets, optimal window size and orientation, and training sample size optimization an open question[12, 25]. Major findings in the literature include the following:

- Harmonic feature combinations without overfitting results in high performance of supervised ML algorithms[12, 25].

- BoW has good performance for disambiguation and simple implementation compared to other single features[25].
- Wider window sizes (entire abstract) surrounding the ambiguous target word provide better performance for WSD within biomedical text[25, 39] compared to general English text[48].
- UMLS CUI as a feature has better accuracy than UMLS semantic type information[39].

To obtain optimal "learning", supervised ML algorithms are required to have enough training samples. Liu et al.[48] found supervised classifiers require at least "a few dozens of instances" for each sense. Xu et al.[51] scrutinized "required sense size," and found that increasing the training sample size tends to diminish the error rate if senses are well separated semantically. They also found that a well-separated sense distribution did not affect performance and error rate corresponds to the similarity of senses, and the major classifier performs competitively if the distribution of the majority sense is more than 90%.

3.3 Methods

3.3.1 Data sets

Clinical notes from Fairview Health Services 2004 to 2008 from four metropolitan hospitals in the Twin Cities were used from our research repository. These 604,944 notes were created primarily from voice dictation and transcription with the option of manual

editing and included admission notes, inpatient consult notes, operative notes, and discharge summaries.

The 440 most frequently used clinical acronyms and abbreviations were identified using a hybrid heuristic rule-based and statistically-based technique. Potential acronyms and abbreviations were chosen if they consisted of capital letters with or without numbers and symbols (periods, comma, colon, or semicolon) and occurred over 500 times in the corpus. For each acronym or abbreviation, 500 random occurrences of the acronym and abbreviation were selected within the corpus, along with the surrounding previous and subsequent 12 word tokens and presented to two physicians to manually annotate for the senses of the potential acronyms or abbreviations. These 500 occurrences could potentially be extracted from the same discourse if the target acronym or abbreviation was repeated within the discourse. We selected 24 surrounding words as a conservative set of surrounding text, since previous work has demonstrated that humans can properly comprehend meaning given approximately five words including an acronym or abbreviation in the center position[10]. The inter-annotator agreement of the annotated sense was reported as Kappa with an overlap of 11,000 instances. Percentage agreement was 92.40% and Kappa statistic was 0.84 overall indicating a reasonable inter-rater agreement. These manual annotations were used as the gold standard.

Among 440 data sets, 50 acronyms and abbreviations were used for this study. We considered those acronyms and abbreviations with a majority sense less than or equal to 95%, then selected the same number of sets according to their majority sense ratio. Table 3.1 shows the 50 acronyms and abbreviations according to their major dominant sense

rates. Table 3.2 summarizes the senses of acronyms and abbreviations and their coverage in the 500 samples. For example, 'CVA' has two different senses "cerebrovascular accident" (278 samples, 55.6% - majority sense) and "costovertebral angle" (222 samples, 44.4%).

Table 3.1 Distributions of annotated senses of selected clinical acronyms and abbreviations

Rate of majority sense	Number	Acronyms and abbreviations
90 – 95%	5	BAL, CVS, DIP, IM, OTC
85 – 90%	5	C&S, CEA, CVP, ER, FISH
80 – 85%	5	ASA, MSSA, PE, SBP, T4
75 – 80%	6	AVR, CA, CTA, IR, NAD, RA
70 – 75%	4	AV, PDA, SA, SMA
65 – 70%	5	AB, BK, DT, LE, RT
60 – 65%	3	IVF, MR, OP
55 – 60%	5	CVA, DC, DM, PCP, VBG
50 – 55%	5	C4, CDI, PAC, PR, T3
45 – 50%	2	C3, T2
Less than 45%	5	AC, IT, MP, PA, T1

Table 3.2 Annotated senses for selected acronyms and abbreviations in clinical corpus

Abbr	Sense	Rate%	Abbr	Sense	Rate%	Abbr	Sense	Rate%
AB	abortion	69.0	DIP	distal interphalangeal	92.4	PA	posterior-anterior	42.4
	blood group in ABO system	27.4		desquamative interstitial pneumonia	7.2		pulmonary artery	27.6
	other 10 senses	3.6		dipropionate	0.4		physician associates	16.6
AC	acromioclavicular	31.8	DM	dextromethorphan	57.2		physician assistant	12.2
	adriamycin cyclophosphamide	23.6		diabetes mellitus	41.8		other 4 senses	1.2
	(drug) AC	21.8		other 2 senses	1.0	PAC	premature atrial contraction	55.0
	before meals	18.8	DT	diphtheria-tetanus	67.2		physician assistant certification	27.4
	other 6 senses	4.0		delirium tremens	25.8		post anesthesia care	9.2
ASA	acetylsalicylic acid	80.8		dorsalis pedis:DP*	4.4		picture archiving communication	5.0
	American Society of Anesthesiologists	18.6		other 4 senses	2.6		other 5 senses	3.4
	aminosalicylic acid	0.6	ER	emergency room	89.6	PCP	pneumocystis carinii pneumonia	58.8
AV	atrioventricular	74.8		extended release	6.8		primary care physician	22.2
	arteriovenous	23.2		estrogen receptor	3.6		phencyclidine	18.6
	other 2 senses	2.0	FISH	fluorescent in situ hybridization	89.8		other 2 senses	0.4
AVR	aortic valve replacement	76.2		General English ('fish')	10.2	PDA	posterior descending artery	72.2
	augmented voltage right arm	20.6	IM	intramuscular	92.2		patent ductus arteriosus	27.6
	other 5 senses	3.2		intramedullary	7.6		patient-controlled analgesia:PCA†	0.2
BAL	bronchoalveolar lavage	91.2		unsure sense	0.2	PE	pulmonary embolus	81.6
	blood alcohol level	8.6	IR	interventional radiology	78.8		pressure equalization	17.8
	unsure sense	0.2		immediate-release	20.4		other 2 senses	0.6
BK	BK (virus)	68.6		other 3 senses	0.8	PR	pr interval	50.4
	below knee	31.4	IT	General English	45.0		per rectum	28.2
C&S	conjunctivae and sclerae	86.8		information technology	20.6		progesterone receptor	17.6
	culture and sensitivity	9.4		intrathecal	11.6		other 3 senses	3.8
	other 3 senses	3.8		ischial tuberosity	9.6	RA	right atrium	78.8
C3	cervical 3	49.8		iliotibial	7.0		rheumatoid arthritis	13.2
	component 3	48.6		intertrochanteric	2.8		room air	7.2
	other 2 senses	1.6		other 4 senses	3.4		other 2 senses	0.8
C4	cervical 4	52.2	IVF	in vitro fertilization	61.6	RT	radiation therapy	67.2
	component 4	46.2		intravenous fluid	37.2		respiratory therapy	29.6
	other 3 senses	1.6		unsure senses	1.2		other 5 senses	3.2
CA	cancer	78.2	LE	leukocyte esterase	68.4	SA	slow acting/sustained action	74.0
	carbohydrate antigen	21.0		lower extremity	26.8		sinuatrial	17.6
	other 2 senses	0.8		other 7 senses	4.8		unsure senses	6.6
CDI	Children's Depression Inventory	54.0	MP	metacarpophalangeal	35.4		other 4 senses	1.8
	center for diagnostic imaging	45.0		mercaptopurine	21.4	SBP	spontaneous bacterial peritonitis	83.4
	other 2 senses	1.0		metatarsophalangeal/metacarpophalangeal	21.0		systolic blood pressure	16.6
CEA	carcinoembryonic antigen	88.6		metatarsophalangeal	10.8	SMA	superior mesenteric artery	70.6
	carotid endarterectomy	10.6		unsure senses	6.8		sequential multiple autoanalyzer	16.8
	other 3 senses	0.8		other 4 senses	4.6		spinal muscular atrophy	11.2
CTA	clear to auscultation	79.2	MR	magnetic resonance	62.8		other 3 senses	1.4
	computed tomographic angiography	20.0		mitral regurgitation	35.2	T1	tumor stage 1	39.6
	other 3 senses	0.8		other 4 senses	2.0		thoracic vertebra 1	38.8
CVA	cerebrovascular accident	55.6	MSSA	modified selective severity assessment	83.6		T1 (MRI)	20.6
	costovertebral angle	44.4		methicillin-susceptible Staphylococcus aureus	16.4		other 2 senses	1.0
CVP	central venous pressure	87.2	NAD	no acute distress	75.4	T2	T2 (MRI)	45.4
	cyclophosphamide, vincristine, prednisone	12.4		nothing abnormal detected	24.6		tumor stage 2	33.2
	cardiovascular pulmonary	0.4	OP	oropharynx	61.6		thoracic vertebra 2	19.4
CVS	chorionic villus sampling	91.4		oblique presentation/occiput posterior	24.2		other 3 senses	2.0
	cardiovascular system	8.2		operative	11.0	T3	triiodothyronine	53.6
	customer, value, service	0.4		other 5 senses	3.2		tumor stage 3	31.2
DC	discontinue	56.4	OTC	over the counter	93.8		thoracic vertebra 3	12.8
	direct current	30.4		ornithine transcarbamoylase	6.2		other 2 senses	2.4
	discharge	6.2				T4	thyroxine	84.8
	District of Columbia	6.2					thoracic vertebra 4	8.2
	other 3 senses	0.8					tumor stage 4	7.0
* DP (dorsalis pedis) should be used instead of DT † PCA (patient-controlled analgesia) should be used instead of PDA						VBG	vertical banded gastroplasty	59.8
							venous blood gas	40.2

3.3.2 Features

For this study, the following features were included and defined as follows:

- Window size is the number of word tokens on each side of the given acronym or abbreviation. Window size was varied as follows: ±3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 words, entire section, and entire document levels. The sentence level was not analyzed separately in our experiments because of the lack of formal sentence structure within clinical notes. On average, the number of word tokens per document section was 67.97 and for a given note was 391. The window size of 3 means that three previous word tokens, the given acronym or abbreviation, and three post word tokens were included in the window. Windows were also examined asymmetrically (e,g., more context on the left of the acronym than on the right) to understand the relative value of the left and right-sided information.
- Bag-of-words (BoW) uses each unique word as a feature in a non-weighted vector, not considering word order. Taking into account frequency and form (i.e. stems) of words, Lexical Variant Generation (LVG)[63] normalization tool distributed with MetaMap was used to normalize the 1,000 most frequent words. We limited normalization to the most frequent items in order to speed up the processing for a large number of experiments conducted in this study. We recognize, however, that normalization of lower frequency words may be of further but likely marginal benefit. We also experimented with BoW both with and without stop words[47] to further reduce the feature space.
- Concept Unique Identifiers (CUIs) were generated from MetaMap. Unique or multiple CUIs were obtained by putting the phrase of a given window size

including the acronym or abbreviation into MetaMap. Metamap also generates a score for each potential mapped CUI with a maximum score of 1000 (high likelihood of a positive match). Score cutoffs were varied in our analysis as shown in the Results section.

- Semantic types were generated from each of the CUI mappings. The feature set consisted of unique or multiple semantic types generated by putting the selected phrase within a given window size including the acronym or abbreviation into MetaMap. Semantic type groups were also used, aggregating into the pre-defined 15 groups proposed by McCray et al.[29]
- Position information in clinical notes was defined as the relative position of the acronym at the section level and document level. Positions were calculated relatively as the location of the target abbreviations over total words of each level.
- Section information from clinical notes is a local contextual feature. We extracted the relevant section information for the given sample as the closest previous section header to the target acronym or abbreviation. Four heuristic conditions were used to detect section information for the given acronym or abbreviation: (1) the previous line is an empty line or other line return symbol only; (2) the position of phrase starts at the beginning of a line; (3) the section indicator symbol ":"; and (4) words from the beginning to the section indicator symbol in the line are written in upper case characters. A physician merged sections tags manually because of the variability in expression for sections names in clinical notes.
- Word level POS tags were generated using the Stanford POS tagger. POS tags were

collected by putting the word chunk with a given word window size including the acronym or abbreviation into the Stanford POS tagger.

3.3.3 Algorithms and evaluation

Three fully supervised classification algorithms (Naïve Bayes, Support Vector Machines, and Decision Tree) were implemented with different window sizes using the 10-fold cross-validation setting in Weka (NaiveBayes, LibSVM, and J48 with the default settings), respectively. Window sizes and orientations were also varied to include different numbers of left or right word tokens to find optimal window orientation. Accuracy was reported for system performance with 10-fold cross-validation. Baseline performance was considered to be the majority sense, which helped in evaluating the performance of our ML algorithms. BoW without LVG or stopwords was used for these simulations as a representative baseline methodology.

Table 3.3 Setting parameters of various cross-validation per acronym or abbreviation

	Inverted cross-validation							Cross-validation		
	100	50	25	20	10	5	4	2	5	10
Number of training samples	5	10	20	25	50	100	125	250	400	450
Number of testing samples	495	490	480	475	450	400	375	250	100	50
Number of simulations	100	50	25	20	10	5	4	2	5	10

To explore minimum training sample sizes for acronyms and abbreviations we used inverted cross-validation (ICV). With IVC, various size sub-sets of samples of the acronym or abbreviation were used one time for testing by ICV and the results for sub-

sets were averaged to assess performance. ICV is a useful approach for estimating the minimum number of samples required to reach stable performance at a desired accuracy level. Because the average number of senses of selected acronyms and abbreviations was 4.72, we used ICV with 100 and lower number of iterations. Table 3.3 illustrates training and testing sample sizes with various ICV or cross-validation for each evaluation. For inverted cross-validation, the average accuracy of simulations was reported for the system performance.

3.4 Results

When aggregating the performance, particularly overall accuracy for 50 acronyms and abbreviations, there were several general findings. From the perspective of classifiers, similar performance was achieved regardless of the classifier type with 10-fold cross-validation. However, SVM classifiers tended to show slightly better performance compared to NB classifiers, and NB classifiers tended to show better performances compared to DT classifiers. With respect to individual features, most features contain better information relative to the baseline majority sense. Among them, BoW features showed better but not statistically different performance compared to other features. As a second best feature, CUI demonstrated better performance than UMLS semantic type with grouping when using the threshold score 900 from MetaMap for a match compared to 1,000.

Increasing window size was found to have a tendency to improve performance at the lower but not the higher end of the window size range. Moreover, entire section and

document-size windows showed further deterioration in performance. In contrast, larger window sizes for POS tag features tended to initially decrease performance at lower sizes and then increase performance at larger sizes. The best window size for classifier performance was found to vary with individual features and classifiers. Using SVM classifiers, the best window size with a symmetric window for BoW was 40 (left 40 and right 40 words) and for CUI features with MetaMap was 45 words. Taking out simple English stopwords resulted in better performance when the window size was larger than 20 words in our dataset using NB classifiers. However, removal of stopwords was not helpful for symmetric windows smaller than 20.

As a single feature, section information alone resulted only in an accuracy of 80%. However, it contributed additional information to other single or combined features. Compared to CUI or semantic type features, the combination of sections with CUI or semantic type features improved the ML performance. Although the combination of sections with BoW features did not perform significantly better than BoW features, this combination still gave enough information to make it the best combination of features from the feature types examined.

Because BoW resulted in best performance, we investigated information contained in each window of BoW using only one side of window. We utilized BoW along with the SVM machine learning algorithms.

Figure 3.1 contains a graphical representation of performance with a symmetric window and with windows containing only words on the right or left side. The figure shows that the left word window of the target acronym or abbreviation contains more information for WSD compared to the right word window. The use of both sides of word windows offers better discriminating information than the left side alone.

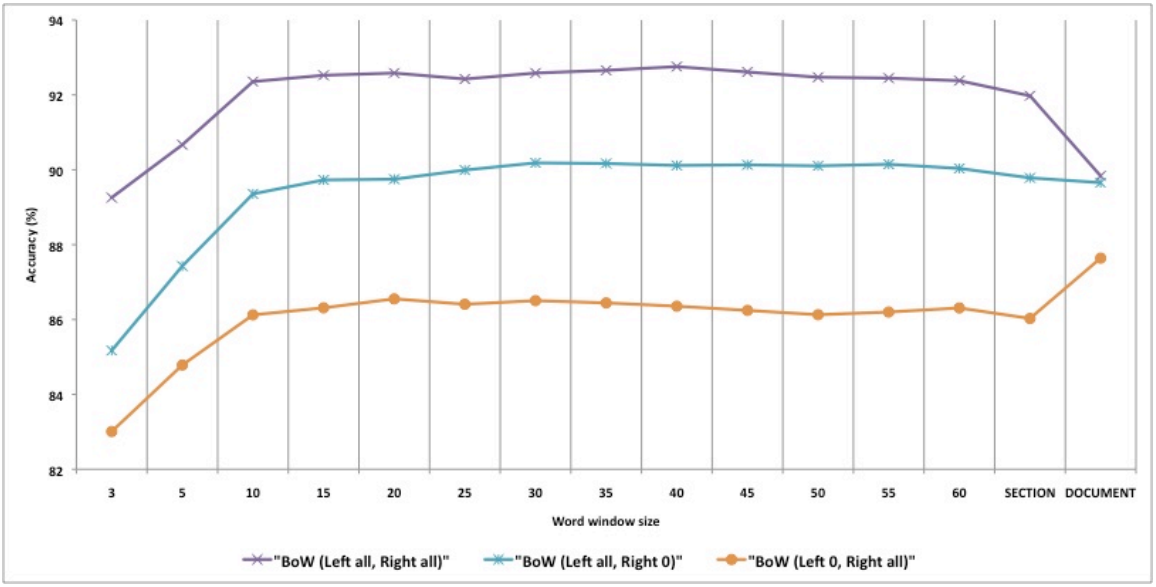

Figure 3.1 Accuracy depending on different sides of word window for BoW with SVM classifiers

The summarized result using SVM classifiers with an expanding left window BoW is shown in Table 3.4 with acronyms and abbreviations separated by majority sense ratios. Table 3.4 illustrates a tendency of acronyms and abbreviations with low majority ratios to require a wider left window for best performance. However, if the majority ratio of acronyms and abbreviations is higher (over 80%), it paradoxically performed best with the entire document (left of the target acronym or abbreviation). When this was repeated with the right window, we observed that the maximum performance with the right window was achieved with the use of the entire right document window regardless of the majority sense ratio.

Table 3.4 Depending on left word window, sub-aggregated accuracies of grouping by majority sense ratios of abbreviations

Left BoW using SVM	<0.5 (7 acronyms)	0.5<&<0.6 (10 acronyms)	0.6<&<0.7 (8 acronyms)	0.7<&<0.8 (10 acronyms)	0.8<&<0.9 (10 acronyms)	0.9<&<0.95 (5 acronyms)
3	71.49	81.94	85.75	86.06	92.26	93.92
5	73.51	85.76	88.48	88.28	93.76	94.20
10	77.17	88.28	90.88	89.54	94.44	95.60
15	77.09	89.30	91.03	90.32	94.52	95.44
20	76.51	89.44	90.90	90.48	94.76	95.56
25	76.94	89.82	91.43	90.48	94.70	95.92
30	77.00	90.00	91.40	91.14	94.68	96.16
35	77.06	89.94	91.75	91.32	94.12	96.24
40	76.94	89.80	91.80	90.90	94.38	96.40
45	77.09	89.90	91.43	90.82	94.64	96.40
50	76.89	90.22	91.48	90.86	94.44	96.00
55	77.43	90.00	91.60	90.78	94.38	96.20
60	76.94	90.00	91.38	90.90	94.18	96.28
Section	76.91	89.40	90.75	90.50	94.36	96.44
Document	77.14	88.84	90.58	89.04	95.04	97.80

Figure 3.2 shows accuracy trends for a fixed left window size (40) and an increasing right word window size (X-axis) with different majority sense distributions. In general, good performance is reached with a smaller size right window. The best performance of BoW is 92.88% (over all 50 acronyms and abbreviations) with 40 left side window and 23 right side window.

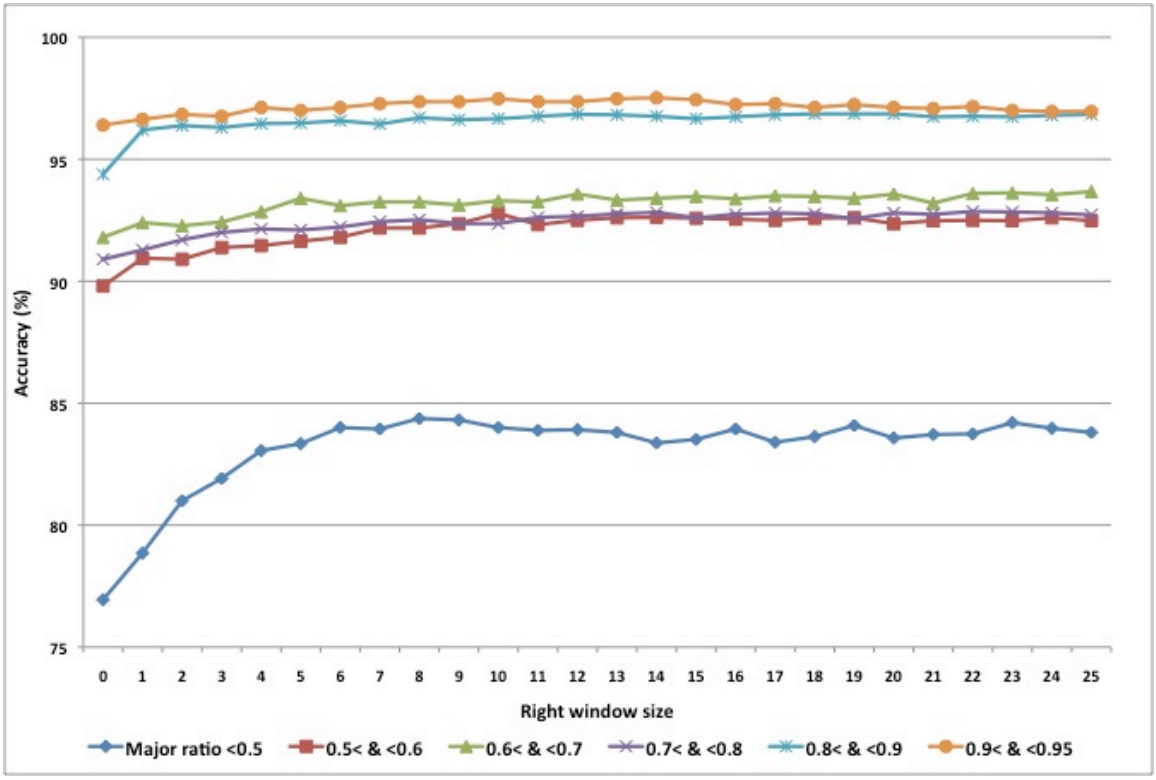

Figure 3.2 Accuracy depending on varying right word window with left 40 word window (Majority ratio = majority sense ration in groups of acronyms and abbreviations)

Figure 3.3 is the aggregated accuracy of 50 abbreviations with both sides of the word windows equal to 40 when using SVM classifiers with BoW with various inverted or standard cross-validation settings. Increasing the training sample size increases the

accuracy for disambiguation as expected. Our findings demonstrate that 2, 5 and 10-fold cross-validations show similar performance. Furthermore, increasing the sample size with ICV shows increasing performance when comparing the gradated performance between 100 ICV and 4 ICV. As shown in Figure 3.3, for a desired accuracy to over 90%, the minimum sample number is 125 (4 ICV) when using SVM classifiers, and approximately 250 (2 CV) when using NB classifiers over the aggregated 50 acronyms and abbreviations. Therefore, when there is little information about majority sense distributions of acronyms and abbreviations, at least 125 training samples is a reasonable baseline required for acronym and abbreviation WSD classification with the SVM classifier.

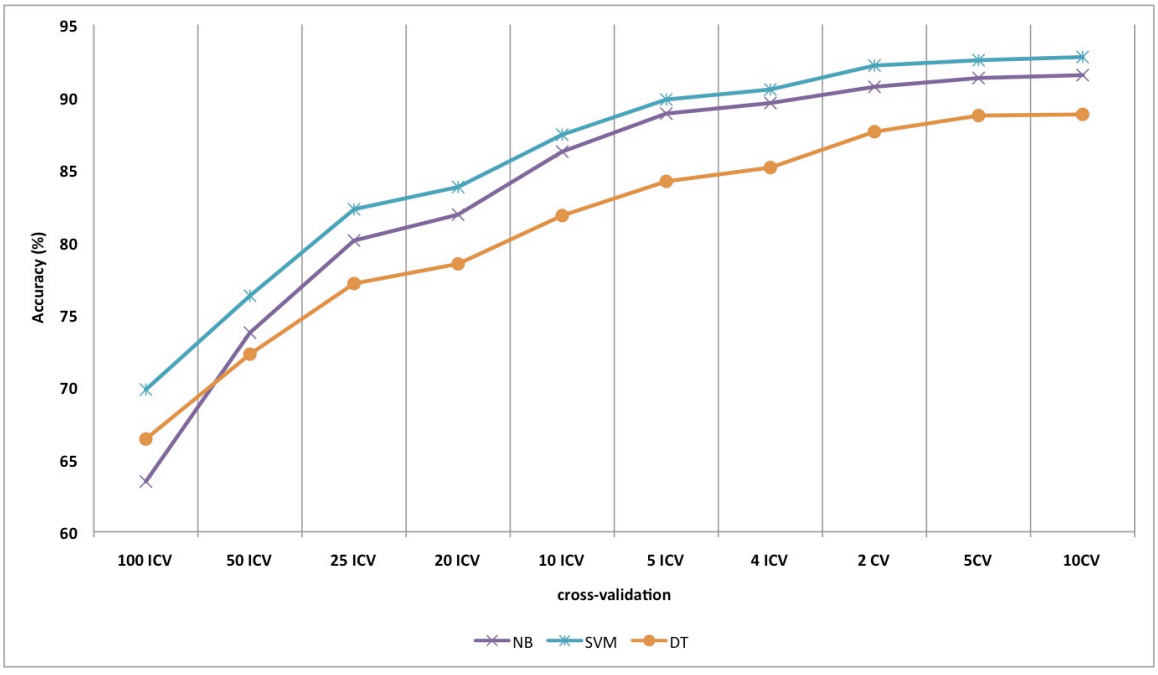

Figure 3.3 Accuracy depending on CV (size of training sample)

Table 3.5 Comparision among classifiers split by majority sense ratio using NB and SVM

Majority ratio	SVM classifier									
	Inverted cross-validation							Cross-validation		
	100	50	25	20	10	5	4	2	5	10
<0.50 (7 acronyms)	42.33	53.31	64.79	67.75	73.99	77.94	79.63	82.66	82.74	83.94
0.50< & <0.60 (10 acronyms)	60.07	70.76	80.27	82.83	86.76	89.36	89.98	91.12	92.84	92.80
0.60< & <0.70 (8 acronyms)	68.34	75.77	83.24	84.66	88.83	91.20	91.58	93.03	93.15	93.28
0.70< & <0.80 (10 acronyms)	74.80	79.34	83.38	84.44	87.86	90.12	90.37	92.40	92.14	92.26
0.80< & <0.90 (10 acronyms)	84.00	87.16	89.23	89.99	92.57	94.63	95.31	96.36	96.74	96.84
0.90< & <0.95 (5 acronyms)	91.72	92.44	92.90	93.02	93.88	94.99	95.83	97.36	97.04	97.00

Majority ratio	NB classifier									
	Inverted cross-validation							Cross-validation		
	100	50	25	20	10	5	4	2	5	10
<0.50 (7 acronyms)	39.23	52.09	62.64	65.48	72.17	76.17	78.01	79.83	80.29	81.17
0.50< & <0.60 (10 acronyms)	53.52	69.17	78.17	80.94	85.56	88.62	89.24	90.14	92.12	92.20
0.60< & <0.70 (8 acronyms)	60.35	71.83	79.55	81.53	87.15	89.68	90.28	92.35	92.83	93.15
0.70< & <0.80 (10 acronyms)	67.23	76.02	81.35	82.92	87.35	89.96	90.25	91.62	92.06	92.12
0.80< & <0.90 (10 acronyms)	79.91	85.07	87.88	88.64	91.28	93.42	93.95	94.96	95.14	95.22
0.90< & <0.95 (5 acronyms)	81.83	88.99	91.12	91.72	93.45	94.43	95.27	94.04	93.52	93.32

Table 3.5 summarizes the accuracy of SVM and NB when grouping acronyms and abbreviations according to the majority sense ratios. The highlighted cells are the first points with over 90% aggregated accuracy across inverse cross validation settings. Here, acronyms and abbreviations with high majority sense ratios tend to require fewer samples than acronyms and abbreviations with low majority sense ratios to achieve a threshold of 90% accuracy. In terms of classifiers, SVM and NB classifiers demonstrated better and more stable performance over the DT classifier. SVM had better performance than the

NB classifier to classify senses of the acronyms and abbreviations when it has fewer samples.

3.5 Discussion

This study provides important insights into the area of clinical acronym and abbreviation WSD. Our main finding is that the left side of words in a window around the target acronym or abbreviation provides better information for disambiguation than the right side of the window. Therefore, an asymmetrical window larger on the left and smaller on the right maintains performance and allows for a smaller feature space and a more efficient computational process. This phenomenon coincides with the process of sense discrimination by human annotators. When annotators classify senses of acronyms and abbreviations, they mainly focus on the left side of target token. Interestingly, humans require a very small number of tokens for the right window (about 5 words) compared to our automated methods (about 20 word window). One factor that could partially account for this discrepancy is that there may be information lost in the pre-processing steps for features (i.e., lexical normalization and selection of 1,000 frequent words). Another main finding of this study was the observation that a size of around 125 samples with SVM classifiers may be effective as a baseline threshold for training. However, it is important to note that in cases of acronyms and abbreviations with less than 50% majority sense ratios, all accuracies were lower than 90% even in 10-fold cross-validation settings, which warrants future study into the enriching datasets with rare sense distributions associated with acronyms and abbreviations with these distribution

patterns.

We extracted the most frequently used 440 acronyms and abbreviations with a cut-off frequency of 500 occurrences from a large corpus consisting of various types of clinical notes and annotated these with experts. To examine questions about training sample size, we carefully selected the acronyms and abbreviations according to the majority sense ratio. While it is possible that these findings are specific to the corpus of text that we used, these results are still helpful to identify representative trends in acronym and abbreviation sense disambiguation in the clinical domain. The large size of the dataset (50 acronyms) is also helpful in elucidating the amount of variability that exists in WSD of acronyms in clinical texts. Some of the parameters are slightly different in these experiments compared to previous studies, several findings from this study on acronym and abbreviation WSD in clinical notes are consistent with several other previous studies[12, 25, 48, 51] of word, acronyms, and abbreviation sense disambiguation in biomedical literature and clinical notes. (i.e., the BoW feature is a powerful feature and SVM algorithm has good performance for WSD). The defining contribution of this work was its use of a large set of clinical acronyms and abbreviations and the examination of both window orientation and size as well as looking at the question about minimum training sample numbers with a systematic approach.

The combination of using all features dropped performance in our results. A possible explanation is the presence of duplicative or conflicting information between different features (especially POS tag feature) with larger window sizes (up to document level). Another possible reason is that CUIs and semantic features may contain noise

from inaccuracies in MetaMap, which was used for CUI mapping. There is also a tendency for clinical texts to contain incomplete sentences and other poorly-formed text. Furthermore, windows for WSD tasks are typically based on centering acronyms and abbreviations in our experience and also sometimes do not maintain full sentences for the Stanford POS tagger or using by MetaMap. As such, the Stanford POS tagger or MetaMap may generate incorrect POS tags or concepts from any partial sentence phrase, which may deteriorate the overall ML performance. Lastly, the Stanford POS tagger may not be optimized for dealing with clinical notes because it is trained and designed for general English.

It is important to note that this is another example where MetaMap may need future optimization as a core of the UMLS. Because the tool was not developed for the clinical domain, it may suffer in performance for certain clinical tasks. According to Savova et al.[12], 20% of pertinent ambiguous terms overlapping between biomedical and clinical domains possess more senses in the clinical domain than the biomedical domain. Xu et al.[3] also found that terms in clinical corpora have low coverage in UMLS. Therefore, we may miss CUI and semantic information in the clinical domains. We also attempted to enhance semantic information by adding semantic grouping information of McCray et al. but found that this did not significantly improve the performance because one of the semantic groups dominates (48.8%): “Chemicals & Drugs”. Furthermore, some groups such as “Genes & Molecular Sequences”, “Geographic Areas”, “Occupations”, etc, are proportionally too small (only 0.1%).

Certain limitations are important to note with this study in its interpretation. The

main limitation is that the features utilized here are based on words and are mostly dependent upon one another. In other words, CUI or semantic information from MetaMap contains overlapped information with BoW. Therefore, performance using BoW features shows similar performance using the combination of knowledge features (BoW+CUI+Semantic information). Another issue is that there is no systematic management implemented for the number of features in this study. The average number of features per instance was 849 for BoW, 2,427 for CUI, and 134 for semantic information when we fix the word window size to 40 symmetrically. In other words, MetaMap features may offer insufficient information for the machine to learn compared to BoW features. There is also the important issue of dealing with rare senses, which drop the system performance significantly and require specific methodologies to address adequately. We did not eliminate these rare senses in this experiment in order to reflect the difficulty of this task with clinical notes, and all rare senses, as well as typographical and other errors in the samples were included in this experiment.

Future work is needed to determine if our methods and findings are scalable for other clinical note corpora. We used a heuristic approach to detect section information which may require modification for other documents, as over 25,000 lexically unique section headers were found in this document repository. Finally, although we assumed that there was "one sense per-discourse", this may not apply throughout an entire clinical document[19] when considering section information, which is an issue that we plan to explore further.

3.6 Conclusion

In this study we investigated a large group of clinical acronyms and abbreviations from our clinical notes corpus to understand issues related to practical clinical acronym and abbreviation WSD. Using 50 clinical acronyms and abbreviations with a majority sense < 95%, we found BoW to be an efficient feature set. When looking at window orientation and size, a symmetric window of ~40 words was found to have good performance with the left side of the window providing more valuable information compared to the right side. Our experiments also demonstrate that an SVM classifier with at minimum 125 training samples was needed to achieve at least 90% accuracy for clinical WSD tasks. These findings provide important insight into the application of clinical acronym and abbreviation WSD in clinical NLP system modules.

Chapter 4 AUTOMATED NON-ALPHANUMERIC SYMBOL RESOLUTION IN CLINICAL TEXTS[3]

Sungrim Moon, MS[1], Serguei Pakhomov, PhD[1,2], James Ryan[3], Genevieve B. Melton, MD[1,4]

[1]Institute for Health Informatics, [2]College of Pharmacy, [3]College of Liberal Arts, [4]Department of Surgery

University of Minnesota, Minneapolis, MN

Although clinical texts contain many symbols, relatively little attention has been given to symbol resolution by medical natural language processing (NLP) researchers. Interpreting the meaning of symbols may be viewed as a special case of Word Sense Disambiguation (WSD). One thousand instances of four common non-alphanumeric symbols ('+', '–', '/', and '#') were randomly extracted from a clinical document repository and annotated by experts. The symbols and their surrounding context, in addition to bag-of-Words (BoW), and heuristic rules were evaluated as features for the following classifiers: Naïve Bayes, Support Vector Machine, and Decision Tree, using 10-fold cross-validation. Accuracies for '+', '–', '/', and '#' were 80.11%, 80.22%, 90.44%, and 95.00% respectively, with Naïve Bayes. While symbol context contributed the most, BoW was also helpful for disambiguation of some symbols. Symbol

[3] This research was supported by the American Surgical Association Foundation Fellowship, the University of Minnesota Institute for Health Informatics Seed Grant, and by the National Library of Medicine (#R01 LM009623-01). We would like to thank Fairview Health Services for ongoing support of this research.

disambiguation with supervised techniques can be implemented with reasonable accuracy as a module for medical NLP systems.

4.1 Introduction

Clinicians frequently use a wide range of shorthand expressions to maximize efficient communication in not only expressing linguistic meanings but also in representing medical information[1]. In addition to large numbers of abbreviations and acronyms, a number of symbols are utilized as condensed meaning-bearing units in free-text clinical notes. Like words, acronyms, and abbreviations, these symbols, which consist mostly of non-alphanumeric characters, often have ambiguous senses. Symbol disambiguation may be considered an analogous problem to automatic word sense disambiguation (WSD). Since the antecedent or pre-processing Natural Language Processing (NLP) module can potentially deteriorate the quality of downstream processing functions of automatic NLP systems[64-66], proper resolution of symbols is necessary to ascertain the meaning of symbols and preempt errors in automated medical NLP systems.

Neither the medical NLP nor computational linguistics literature has focused upon symbol resolution to any large extent. In the biomedical domain, researchers have investigated disambiguation of gene symbols from biomedical text. In one such study, gene symbol disambiguation was performed with the goal of identifying biomedical entities[67]. Computational linguists, in contrast, have been mainly interested in the meaning of words themselves and have largely ignored non-alphanumeric symbols

outside of dealing with the task of sentence splitting.

In one analogous study that focused on symbol resolution in Chinese text, Hwang et al. examined resolution of three non-alphanumeric symbols (‘/’, ‘:’, and ‘–’) in the Academic Sinica Balance Corpus (ASBC), which consists of Mandarin and English symbols[68]. They found seven senses for symbol ‘/’, five senses for ‘:’, and seven senses for ‘–’. They set up a rule-based multi-layer decision classifier (MLDC) utilizing applied linguistic knowledge with a statistical voting schema and used words surrounding the target words (bag-of-words, BoW) with statistical probabilities as features. This two-layer model was expanded into a three-layer model using preference scoring based on the location of characters/words[69]. While this approach may be effective in some cases, rule-based classification with linguistic knowledge can serve as a bottleneck in maintaining automatic resolution systems because language is always changing and these rules must be maintained depending on characteristics of the corpus. Even if the MLDC used by these authors focused upon symbol disambiguation, this is at best an analogous application to English clinical note disambiguation. These results may not be directly transferrable to clinical notes because of the structural difference between English and Mandarin, and because of contextual difference between general documents and clinical notes. For example, English word tokens are separated by whitespace, but Mandarin word tokens are not.

For this pilot study, we selected four symbols (‘+’, ‘–’, ‘/’, and ‘#’) and conducted a set of experiments for automated symbol sense disambiguation using clinical notes. We investigated symbol senses using the literature and annotations of a moderate-sized

corpus, and then performed automated symbol disambiguation using three supervised machine-learning classification algorithms: Naïve Bayes, Support Vector Machine, and Decision Tree classifiers).

4.2 Method

4.2.1 Symbol sense inventory

An initial sense inventory for the target symbols ('+', '–', '/', and '#') was created from several reference resources. From the field of computational linguistics, we utilized two textbooks: Speech and Language Processing and Foundations of Statistical Natural Language Processing[47, 70]. We also identified several medical references with symbol senses including a medical dictionary (*Stedman's Medical Abbreviations, Acronyms & Symbols*[55]), medical terminological reference (Medical Terminology and references of approved symbols[19, 26, 71]), and references from the clinical literature (Abbreviations and acronyms in healthcare[5]).

The symbol sense inventory was then refined to remove unclear senses and add missing senses identified by a clinician (GM), and two linguists (JR and SP). "Literature sense" represents this initial sense inventory for the target symbols.

4.2.2 Experimental samples and document corpus

The document corpus for this study consisted of electronic clinical notes from University of Minnesota-affiliated Fairview Health Services (consisting of four metropolitan hospitals in the Twin Cities), containing admission notes, discharge summaries, operative reports, and consultation notes created between 2004 and 2008.

For non-alphanumeric symbols of interest ('+', '–', '/', and '#'), a target instance of a symbol was defined as the presence of the symbol character within a target token. For the purposes of this pilot, the symbols from institution-specific formatting and various section/headers were excluded. For each symbol, 1,000 instances within the corpus were randomly selected for manual annotation.

4.2.3 Reference standard

Using the General Architecture for Text Engineering (GATE) toolkit[72], each of the 1,000 target symbol instances was marked up within each document to clarify and streamline the process of annotating each target symbol. This was particularly important, as multiple instances of potential symbols may exist within a given text or a given target word token. Although studies have demonstrated that most individuals can interpret the proper meaning of a word with a window size of five[7, 10], we provided the entire document during annotation of symbols to ensure adequate context.

Our reference standard was created by two annotators with expertise in medicine and linguistics respectively. Because '+' had several medicine-specific meanings, the annotator for this set was a physician. Since meanings of the other four symbols were less

medically-specific, a linguist (JR) annotated these samples. Whenever the linguist or physician had questions as to the sense of a symbol, these examples were presented and adjudicated with the assistance of two of the authors with linguistics and medical expertise respectively (SP and GM). "Clinical Corpus Sense" represents this empirically-derived clinical sense inventory for the target symbols. Separately, a second annotator examined 200 random samples (50 per symbol) to establish inter-rater reliability of these annotations with percent agreement and Kappa statistic.

4.2.4 Automated system development and evaluation

We created an initial set of features based on the BoW approach to feature extraction and word-form information within the target and surrounding word tokens. These were compared to the majority sense distribution as the baseline. Three fully supervised classification algorithms were applied to these feature sets in a 10 fold cross-validation setting. These algorithms are Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) implemented with NaïveBayes, LibSVM, and J48 using Weka software[73]. We separated 100 random samples from our 1,000 instances of each symbol to determine additional heuristic rules associated with word-form information. After developing the system on 100 random instances, then we evaluated the 900 instances using a 10 fold cross-validation setting on these samples for our result. We report accuracy, recall, precision, and f-measure of our system performance.

4.2.4.1 Basic features

Basic features used as inputs for the three classifiers were:

- Target word token w containing the symbol.
- Prefix and postfix of symbol within the targeted word token w.
- Previous word tokens w-n, target word token w, and post one word token w+n without stemming (BoW with window size n). We explored the optimal window by varying its size and the effect on performance.

In the example: "….erythema. DTRs are diminished at 1+/4+ in the upper and lower extremities….", if the first '+' symbol (bolded) is the target symbol, the target word token w is "1+/4+", the prefix is "1", the postfix is "/4+". BoW with window size 1 is {at, 1+/4+, in} and BoW with window size 2 is {diminished, at, 1+/4+, in, the}.

Beside basic features, we experimented with stop word removal with BoW using a standard list of 57 English stop words[47]. With our previous example, stop word removal with BoW window size 1 is {diminished, 1+/4+, upper}, and the set of BoW with window size 2 and without stop words is {DTRs, diminished, 1+/4+, upper, lower}.

4.2.4.2 Heuristic features

We tested heuristic rules as additional features. Heuristic rules were developed to identify word-form representations of the target word token w or surrounding word tokens (w-n and w+n). 100 random instances from 1,000 were separated for each symbol to develop heuristic rules. Utilizing regular expressions, heuristic rules applied to the target word token w or surrounding word tokens. These were added as additional features

to our classifiers.

Table 4.1 Senses for symbols

Symbol	Literature Sense	Reference	Clinical Corpus Sense
+	acid (reaction)	SMAAS, SHC, ICOIEI	
	added to	SMAAS	
	convex lens	SMAAS	
	decreased or diminished (reflexes)	SMAAS	reflexes
	excess	SMAAS	excess
	less than 50% inhibition of hemolysis (Wassermann)	SMAAS	
	low normal (reflexes)	SMAAS	edema(swelling)
	markedly impaired (pulse)	SMAAS	pulse
	mild (pain or severity)	SMAAS	
	plus	SMAAS, SHC, ICOIEI, Kuhn	plus
	positive (laboratory test)	SMAAS, SHC, ICOIEI	positive (laboratory test)
	present	SMAAS	present
	slight reaction or trace (laboratory tests)	SMAAS	
	sluggish (reflexes)	SMAAS	strength
	somewhat diminished (reflexes)	SMAAS	blood type
	and	ICOIEI, Kuhn	and
			pregnancy dating
			heart murmur
			fetal position during labor
			tonsil size
			uncommon rating
–	line-breaking hyphens	FSNLP	line-breaking hyphens
	lexical hyphens	FSNLP	lexical hyphens
	compound pre-modifiers	FSNLP	compound pre-modifier
	quotative or expressing a quantity or rate	FSNLP	quotative or expressing a quantity or rate
	typographic conventions	FSNLP	typographic convention
	phone number	FSNLP, SLP	phone number
	minus	SHC, ICOIEI	minus
			date
			negative
			and
			and(fraction)
			compound
			hyphenated name
			junction
			obstetrical data
			protocol number
			to
			ZIP+4 code
/	divided by	SMAAS	divided by
	either meaning	SMAAS	either meaning
	extension	SMAAS	
	extensors fraction	SMAAS	
	of	SMAAS	of
	per	SMAAS, Kuhn	per
	to	SMAAS	
	date	SLP	date
	separates two doses	Kuhn	separates two doses
			over(e.g., blood pressure)
			abbreviation
			phone number
			respectively
#	fracture	SMAAS	
	gauge	SMAAS	gauge
	number	SMAAS, MT, ICOIEI, FSNLP	number
	pound	SMAAS, MT, ICOIEI	
	weight	SMAAS	
			quantity
			level

SMAAS = Stedman's Medical Abbreviations, Acronyms & Symbols (Forth Edition)
SHC = Stanford Hospital and Clinics approved abbreviations acronyms and symbols
ICOIEI = Illinois College of Optometry and Illinois Eye Institute
Kuhn = Abbreviations and acronyms in healthcare: When shorter isn't sweeter
MT = Medical Terminology the language of health care second edition
SLP = Speech and Language Processing
FSNLP = Foundations of Statistical Natural Language Processing

Table 4.2 Definition, examples and numbers of symbol senses in clinical documents

Symbol	Clinical Corpus Sense	Definition	Example	N†
+	pulse	used in pulse degree format	pulses are 2 + bilaterally	287
	edema(swelling)	used in edema degree format	4 + brawny edema	187
	reflexes	used in reflexes degree format	2 + patellar reflexes	148
	pregnancy dating	using in pregnancy dating format	38 + 3 weeks' gestation	115
	excess	more than the given number	20 + years, 37 + weeks	68
	strength	used in strength degree format	strength of the upper extremities is 5+	52
	plus	addition between two numbers	49 + 5 cm	35
	heart murmur	used in heart murmur degree format	there was + 1 mitral regurgitation	23
	blood type	indicates antigen to blood type	a blood type A +	21
	positive (laboratory test)	react to laboratory test	blood pressures with 1-2 + protein	18
	uncommon rating*	uncommon rating	left knee has a 2+ effusion	15
	and	functions like the conjunction *and*	caltrate 600 + vitamin D1	11
	present	exist or react	+/+ trigger points	11
	fetal position during labor	position format during labor	the cervix at + 1 to +2 station	6
	tonsil size	indicates of size of tonsil	3 + tonsils	3
–	quotative, or expressing a quantity or rate	appears in quotatives, or constructions expressing a quantity or rate	5-years-old, once-in-a-lifetime	252
	compound pre-modifier	appears in compound pre-modifiers	seizure-like symptoms	226
	compound	links components of a non-modifier compound	K-Dur, x-ray, E-coli, break-through	157
	lexical hyphen	links small word formatives and content words	non-medically, ex-smoker	126
	to	indicates a range	3-4 times	111
	typographic convention	typographic-conventional hyphen or dash	allergies – none.	54
	junction	notes the junction of two elements, usually vertebrae	status post C3-C4 laminectomy	24
	phone number	used in phone-number formatting	612-555-5555	13
	and (fraction)	links an integer and fraction to form a non-integer number	37-1/2 weeks gestation	10
	obstetrical data	appears in what is usually four-pronged data about a patient's pregnancy history	para 0-0-1-0	7
	hyphenated name	links two components of a hyphenated name, usually a surname	Avera-McKennan Hospital	7
	and	functions like the conjunction *and*	type II-III odontoid fracture	3
	date	used in date formatting	05-17-2003	3
	negative	indicates a negative number	-2.132	2
	line-breaking hyphen	follows the first portion of a word that is split by a line break	postopera- tively	2
	ZIP+4 code	separates a zip code and ZIP+4 code	55433-5841	1
	protocol number	serves specification function in an institution's protocol-numbering system	per our protocol #2005-02	1
	minus	indicates subtraction operation	normal 24 + or - 3 ml/kg	1
/	date	used in date formatting	05/17/2003	499
	over(e.g., blood pressure)	couples systolic and diastolic blood pressure measurements, or inhalation and exhalation with BiPAP settings	blood pressure 140/90, we will continue BiPAP at 10/5	196
	either meaning	used in constructions indicating either/both words	and/or, DNR/DNI, Heme/Onc	119
	of	separates a specific rating and the maximum value possible given the scale	regular rate and rhythm with a 2/6 systolic murmur	60
	separates two doses	indicates two separate dosages, usually in drugs with multiple drug constituents	advair 250/50	43
	divided by	separates the numerator and denominator in a fraction	1/2 day, 3-5/7 weeks	39
	per	shorthand for *per*	mg/dL	30
	abbreviation	used to abbreviate, or to link components of an acronym	OB/GYN	6
	respectively	couples values that are each respective to a distinct measure	DP and PT are 1+/4+	6
	phone number	used in phone-number formatting	612/123-4567	2
#	number	shorthand for *number*	hospital day #2	856
	quantity	indicates a quantity, usually of pills dispensed	#10 tablets, #20 dispensed	130
	gauge	indicates gauge specification	aortic valve replacement with #23 medtronic Mosaic valve	13
	level	indicates what level a measurement is at	hemoglobin at #10	1

† N = the number of samples per sense of given symbol in 1000 random samples
* Uncommon rating = subspecialty or other uncommon standard rating

4.3 Results

Table 4.1 compares literature senses from reference sources and the experimental clinical corpus senses in our repository for each symbol. This comparison is organized in the alphabetical order of literature senses. Depending upon the domain, a different set of senses was identified. Table 4.2 depicts the sense, its definition, an example, and the distribution of senses within the corpus. Table 4.2 is ordered based on the sense distribution of each symbol within the clinical corpus. When developing our module, we introduced heuristic rules for this pilot as depicted in Table 4.3.

Table 4.3 Heuristic rules used as additional features to classifier

Symbol	Regular expression	Description of form	Applied sense
+	m/^[1-3]\+/	1+, 2+, 3+	pulse, edema, reflex, excess
	m/^\+[1-3]/	+1, +2, +3	pulse, edema, reflex, excess
	m/^[1-9]?[0-9]\+[0-9]\W?$/	one or two digits for weeks with both side of '+'	pregnancy dating
	m/^[1-9][0-9]\+\W?$/	two digits for years/weeks with previous side of '+'	excess
–	m/^[1-9][0-9]\-/	two digits with previous side of '–'	
	m/^.+\-[1-9][0-9]$/	two digits with post side of '–'	
	m/^[a-zA-Z]?\-[a-zA-Z]?$/	two alphabetic words with both side of '–'	compound, lexical hyphen
	m/^[a-zA-Z]?\-[a-zA-Z]\-[a-zA-Z]?$/	three alphabetic words with both side of two '–'	quotative
/	m/^(1[0-9][0-9])\/((1?[0-9][0-9])\W?)$/)	two or three digits with both side of '/'	over(e.g., blood pressure)
	m/^([a-zA-Z]+)\/([a-zA-Z]+)\W?$/	two alphabetic words with both side of '/'	either meaning
	m/^[0-9]\/([0-9])\W?$/	two digits with both side of '/'	of
#	m/^\#1[0-5]\W*\.*\W*$/ or m/^\#[1-9]\W*\.*\W*$/)	one or two digits for days with post side of '#'	number
	m/^\#[1-4][0|5]\W*\.*\W*$/	two digits for quantity with post side of '#'	quantity

Table 4.4 Frequency in total corpus and inter-rate agreement of symbols

Symbol	Frequency	Proportion agreement (%)	Kappa statistic
+	118,283	100	1.00
–	4,821,029	88	0.86
/	4,785,691	96	0.95
#	721,655	90	0.72

Within the overall corpus of 604,944 notes, the frequency of '+', '–', '/', and '#' are represented in Table 4.4. For inter-rater reliability, 50 random samples were annotated by a second annotator. Proportion agreement and Kappa statistic of each symbol in Table 4.4 indicates respectively reasonable inter-rater agreement even if it is conducted in a small size of samples.

Table 4.5 Performance of Naive Bayes, Support Vector Machine, and Decision Tree classifiers

Symbol	Feature	Naïve Bayes				Support Vector Machine				Decision Tree			
		Acc*	Pre*	Sen*	F-m*	Acc*	Pre*	Sen*	F-m*	Acc*	Pre*	Sen*	F-m*
+	Majority	0.29	0.08	0.29	0.13	0.29	0.08	0.29	0.13	0.29	0.08	0.29	0.13
	Target token	0.47	0.52	0.47	0.41	0.52	0.47	0.52	0.47	0.49	0.49	0.49	0.45
	Target token, Prefix/postfix	0.54	0.51	0.54	0.48	0.54	0.50	0.54	0.48	0.49	0.49	0.49	0.45
	Target token, BoW (size = 1)	0.68	0.66	0.68	0.63	0.41	0.56	0.41	0.36	0.65	0.67	0.65	0.62
	Target token, BoW (size = 2)	0.77	0.74	0.77	0.74	0.32	0.57	0.32	0.20	0.66	0.67	0.66	0.63
	Target token, BoW (size = 3)	0.79	0.75	0.79	0.76	0.30	0.37	0.30	0.15	0.65	0.67	0.65	0.63
	Target token, BoW (size = 4)	**0.80**	**0.78**	**0.80**	**0.78**	0.29	0.22	0.29	0.13	0.65	0.67	0.65	0.62
	Target token, BoW (size = 5)	0.80	0.78	0.80	0.78	0.29	0.22	0.29	0.13	0.65	0.67	0.65	0.63
–	Majority	0.25	0.06	0.25	0.10	0.25	0.06	0.25	0.10	0.25	0.06	0.25	0.10
	Target token	0.62	0.73	0.62	0.61	0.63	0.68	0.63	0.62	0.63	0.79	0.63	0.63
	Target token, Prefix/postfix	**0.80**	**0.80**	**0.80**	**0.79**	0.66	0.79	0.66	0.67	0.63	0.79	0.63	0.63
	Target token, BoW (size = 1), Prefix/postfix	0.77	0.77	0.77	0.76	0.32	0.65	0.32	0.24	0.63	0.79	0.63	0.63
/	Majority	0.51	0.26	0.51	0.34	0.51	0.26	0.51	0.34	0.51	0.26	0.51	0.34
	Target token	0.57	0.49	0.57	0.46	0.61	0.64	0.61	0.55	0.51	0.26	0.51	0.34
	Target token, Prefix/postfix	0.84	0.86	0.84	0.83	0.64	0.75	0.64	0.57	0.75	0.81	0.75	0.71
	Target token, BoW (size = 1), Prefix/postfix	**0.90**	**0.89**	**0.90**	**0.90**	0.54	0.60	0.54	0.40	0.72	0.66	0.72	0.62
	Target token, BoW (size = 2), Prefix/postfix	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.72	0.66	0.72	0.66
#	Majority	0.85	0.72	0.85	0.78	0.85	0.72	0.85	0.78	0.85	0.72	0.85	0.78
	Target token	**0.95**	**0.94**	**0.95**	**0.94**	0.94	0.93	0.94	0.93	0.95	0.94	0.95	0.94
	Target token, Prefix/postfix	0.92	0.92	0.92	0.92	0.94	0.92	0.94	0.93	**0.96**	**0.94**	**0.96**	**0.95**
	Target token, BoW (size = 1)	0.94	0.95	0.94	0.95	0.86	0.86	0.86	0.80	0.95	0.95	0.95	0.94

Acc = Accuracy, Pre = Precision, Sen = Sensitivity, F-m = F-measure

When we applied three supervised machine-learning algorithms with our feature sets, NB classifier had the most stable overall performance compared to both SVM and DT classifier. We tested removal of stop words; however, there was no performance improvement. We also added heuristic rules as described in Table 4.3, but there is no significant change in algorithm performance either. Our results with respect to the

accuracy, recall, precision, and f-measure with the NB, SVM, and DT classifiers using the basic feature set alone and with BoW are summarized in Table 4.5. These results are based on 900 test samples considering all separated senses in Table 4.2. Maximum accuracy for symbol '+' was 80.11%, symbol '–' - 80.22%, symbol '/' - 90.44%, and symbol '#' - 95.00% with the NB classifier. For '+' and '/', using BoW as features provided improved performance with the NB classifier (Table 4.5), but the optimal window size was different for each symbol. For '–', BoW did not contribute additional information for symbol disambiguation. For '#', the target symbol alone was the dominant feature of importance.

4.4 Discussion

We examined non-alphanumeric symbol disambiguation, an under-studied pre-processing NLP function in the clinical domain. To gain a more thorough understanding of symbol sense ambiguity, we performed a survey of the literature and generated an empiric sense inventory, which helped to refine the overall inventory. Symbol disambiguation appears to perform well with simple sets of features but requires different combinations of features for individual symbols. In each case, a relatively small set of features based on the symbol and its context were effective, indicating that this is a relatively simpler task than sense disambiguation for words, acronyms and abbreviations. Despite the relative simplicity of the task, it has been largely ignored in the clinical NLP literature but constitutes an important problem for NLP of clinical documentation. For example, being able to determine the context appropriate meanings of symbols can

contribute to improved named entity recognition and classification.

While the surrounding context, including words beyond the target token, were expected to be important, we found that in the cases of '#' and '–', words beyond the target word w were unnecessary. In fact, for '#', the target word w alone was sufficient for excellent performance. In contrast, senses related to '+' required surrounding context (optimized with window size 4) for optimal performance. One of the main reasons for these differences is that symbol resolution is affected by the number of senses in the sense inventory and proportion of the majority sense of each symbol. In the previous example, the '#' symbol has fewer senses and has higher proportion of the dominant sense (only 4 senses and the majority sense prevalence is 85%) compared to '+' symbol with has 15 possible senses with well-balanced distributions. For '–' symbol, it only required isolated and condensed token information (pre and postfix features) to determine the right meaning. Another potential reason is the degree of semantic relatedness among senses in a given symbol. For example, '#' symbol has 4 senses that are all closely related with the concept 'number'. Thus, disambiguation of the '#' symbol results in better performance compared to '–' symbol, which had a variety of concepts such as 'minus' or several lexical expressions (e.g., lexical hyphens, compound pre-modifier).

We also expected heuristic rules to contribute positively to system performance but found that they were not helpful in our experiment. These rules could be helpful for enumerated items such as dates or telephone numbers where training sets may not be sufficient to capture the large number of possible combinations. We speculate that this did not change performance much since both of these items were low-incidence. Also,

some rules are language and format-specific. For example, the form of date with symbol '–' can be different according to location. The sequence of date, month, and year are opposite in Europe/Asia compared with the United States. Because of these limitations and perhaps some overlap with our general form-based features (thereby not being independent of our heuristic rules), heuristic features did not contribute significantly to system performance.

With the '+' symbol, we discovered that there were a number of senses that were specific to subspecialties or occurred less often, which we combined into a single annotation called "uncommon rating". In contrast, common ratings such as that for edema or reflexes were separated out as separate senses. For example, the sense 'effusion of a joint' (e.g., "Left knee has a 2+ effusion…") or 'prostate size' (e.g., "His is prostate is 1 to 2+ enlarged…") are standard but occur with low frequency. If we group less common senses into a single annotation, the performance of automatic symbol resolution module improves. In this study, we grouped these less common senses into one sense. If we extend this, all kinds of senses such as pulse, strength, reflexes, edema, and uncommon ratings for symbol '+' can be grouped together. As expected, with this aggregate set of senses, the accuracy of NB classifier was 88.89%, up from 81.56% when more common ratings were separated from the less common ratings. Because these sense grouping decisions can be somewhat arbitrary or tailored to the purpose of the NLP module, concrete agreement between annotators and a clear understanding of the goals of the particular symbol disambiguation NLP module's scope are essential.

Another issue is that some senses share the same BoW or the same form of the

target token. In the symbol '–' set, for example, "follow-up", "well-nourished" and "seizure-like" can be a lexical hyphen and/or a compound-premodifier. For example, "He arrived on time for his follow-up" and "His symptoms were seizure-like", these '–' instances are categorized as lexical hyphens, while "We scheduled his follow-up appointment" and "He experienced seizure-like symptoms" are considered to be both lexical hyphens and compound pre-modifier hyphens. These shared forms between senses may create difficulties with disambiguation and may require additional syntactic information such as part-of-speech and syntactic phrase category. The distinction between lexical hyphens and compound pre-modifier hyphens is probably too small to be of practical importance in an NLP system; however, in this exploratory study we chose separate annotations for these entities that may be collapsed.

Our research demonstrates that non-alphanumeric symbol disambiguation is feasible, with good performance on clinical text using standard form-based rules. These rules require some calibration for each symbol type with respect to window size for individual symbols. Since the set of non-alphanumeric symbols is finite (vs. words and acronyms), development of fully supervised disambiguation classifiers is likely to be the most effective and accurate approach. We plan to extend this module to other symbols, including alphabetic symbols, such as 'x', as well as additional non-alphanumeric symbols, with the goal of utilizing these techniques within a pre-processing module for down-stream information extraction functions from clinical text.

4.5 Conclusion

Although symbols, primarily non-alphanumeric characters, are used widely to convey a variety of meanings in clinical discourse, symbol resolution has been less studied by the linguistics and medical NLP communities. Symbol resolution can be viewed as a specific type of WSD, as well as a basic module for automatic medical NLP systems. In this paper, we examined four symbols (‘+’, ‘–’, ‘/’, and ‘#’) to detect clinical symbol senses and to contrast with senses attested in the literature. We found that while supervised machine learning approaches with form-based features to be effective, calibration of features for disambiguation may be needed for system optimization with individual symbols.

Chapter 5 SUMMARY AND FUTURE DIRECTION

This body of work addresses several important challenges with clinical acronym and abbreviation WSD and extends this work to symbol disambiguation by building a sense inventory and exploring symbol WSD in clinical texts. Overall, 1) the comprehensive sense inventory for clinical acronyms and abbreviations is built based on manual annotation and typical biomedical resources, 2) the window size and orientation for an optimal feature set, and the estimated minimum training sample number are investigated, and 3) the feasibility of automated symbol resolution as an extended clinical WSD is demonstrated.

A comprehensive clinical sense inventory for 440 acronyms and abbreviations is mapped via long forms from UMLS, ADAM, and *Stedman's*. This sense inventory addresses a current need and identifies particular characteristics of clinical senses such as skewed sense distributions, practice-specific senses, and incorrect uses. Also, other medical resources cover clinical senses with very low ratios such as the 178 long forms that had no coverage in any biomedical resource even after lexical and semantic mapping processes. With de-identified sentence datasets, this comprehensive clinical sense inventory is the current largest resource for WSD of clinical acronyms and abbreviations.

To achieve efficient automatic sense resolution, optimized window and orientation of feature sets, and estimated size for training using 50 selective clinical acronyms and abbreviations is studied. With a systematic approach, our study shows the left side of the window offers valuable information compared to the right side. Moreover, unlike biomedical literature, around 40 windows of BoW show better ML performance rather

than wider windows such as a section or document window. When considering training ML algorithms, the minimum of 125 training samples are needed to achieve an accuracy of over 90%. These findings provide important clues as to how to improve clinical WSD tasks.

Basic technologies of WSD tools will be directly extended and applied to symbol disambiguation in clinical notes. The symbol senses disambiguation is the first attempt in medical domains, even if determination of appropriate meanings of symbols is an essential task for clinical NLP. The symbol sense inventory (59 senses for four symbols) is built based on manual annotation in clinical notes and compared to senses in various reference resources in computational linguistics, medical literature, and *Stedman's*. In experiments, automated symbol resolution demonstrates the feasibilities (average accuracy is 86.44%) utilizing common feature sets (BoW with word form information) with supervised ML of clinical WSD for acronyms and abbreviations.

Future experiments will focus on deeper topics for WSD tools. To improve the quality of automated WSD tools, refinements such as sense inventory integration[37] with biomedical resources and technologies, generalizations of insightful findings for clinical WSD for acronyms, abbreviations, and symbols utilizing other inter-institutional study, and investigation to demonstrate unsuitability of one-sense per-discourse in clinical WSD domains will be extensively studied.

Finally, additional topics connected with deeper issues for WSD tools are also necessary to achieve the optimal performance for clinical WSD. Identification of optimal feature sets to minimize overlapping information, new or rare sense detection utilizing

semantic information, similarity measurements between senses as a tool for aiding WSD, and utilization of unsupervised ML algorithms instead of supervised ML algorithms should be investigated. Also, increased awareness of safety risk issues surrounding the use of acronyms and abbreviations can be made based upon this work and observations.

Findings and resources created from this thesis will also be useful not only for the medical NLP community but also for the overall medical and biomedical fields. Medical NLP researchers who have had difficulties dealing with a lack of clinical resources can utilize this publically available de-identified sentence dataset of acronyms/abbreviations/symbols to evaluate their algorithms/methods. Although these resources are not designed specifically for this purpose, biomedical researchers or general medical physicians might also use our inventories and datasets as a reference to understand usages of acronyms/abbreviations/symbols within clinical notes.

BIBLIOGRAPHY

1. Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. *Proc AMIA Symp.* 2002:742-746.

2. Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annu Symp Proc.* 2005:589-593.

3. Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. *AMIA Annu Symp Proc.* 2007:821-825.

4. Liu HF, Lussier YA, Friedman C. A study of abbreviations in the UMLS. *J Am Med Inform Assn.* 2001:393-397.

5. Kuhn IF. Abbreviations and acronyms in healthcare: when shorter isn't sweeter. *Pediatr Nurs.* Sep-Oct 2007;33(5):392-398.

6. Walsh KE, Gurwitz JH. Medical abbreviations: writing little and communicating less. *Arch Dis Child.* Oct 2008;93(10):816-817.

7. Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol.* Jun 2005;12(5):554-565.

8. Pakhomov S. Semi-supervised Maximum Entropy based approach to acronym and abbreviation normalization in medical texts. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania: Association for Computational Linguistics; 2002:160-167.

9. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. *Proc AMIA Symp.* 2001:189-193.

10. Kaplan A. An experimental study of ambiguity and context. *Mechanical Translation.* 1950;2(2):39-46.

11. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.* 2008:128-144.

12. Savova GK, Coden AR, Sominsky IL, et al. Word sense disambiguation across two domains: biomedical literature and clinical notes. *J Biomed Inform.* Dec 2008;41(6):1088-1100.

13. Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput.* 2003:451-462.

14. Adar E. SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics.* 2004;20(4):527-533.

15. Wren JD, Garner HR. Heuristics for identification of acronym-definition patterns within text: Towards an automated construction of comprehensive acronym-definition dictionaries. *Method Inform Med.* 2002;41(5):426-434.

16. Ao H, Takagi TI. ALICE: An algorithm to extract abbreviations from MEDLINE. *J Am Med Inform Assn.* Sep-Oct 2005;12(5):576-586.

17. Chang JT, Schutze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE. *J Am Med Inform Assoc.* Nov-Dec 2002;9(6):612-620.

18. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* Mar 1970;48(3):443-453.

19. Xu H, Stetson PD, Friedman C. Methods for building sense inventories of abbreviations in clinical notes. *J Am Med Inform Assoc.* Jan-Feb 2009;16(1):103-108.

20. Liu H, Friedman C. Mining terminological knowledge in large biomedical corpora. *Pac Symp Biocomput.* 2003:415-426.

21. Zhou W, Torvik VI, Smalheiser NR. ADAM: another database of abbreviations in MEDLINE. *Bioinformatics.* 2006;22(22):2813-2818.

22. Liu HF, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assn.* Nov-Dec 2002;9(6):621-636.

23. NIH. Unified Medical Language System (UMLS). 2011; [http://www.nlm.nih.gov/research/umls/].

24. Moon S, Pakhomov S, Melton G. Automated Disambiguation of Acronyms and Abbreviations in Clinical Texts: Window and Training Size Considerations. *AMIA Annu Symp Proc.* 2012:1310-1319.

25. Joshi M, Pakhomov S, Pedersen T, Chute CG. A comparative study of supervised learning as applied to acronym expansion in clinical reports. *AMIA Annu Symp Proc.* 2006:399-403.

26. IHTSDO. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). 2011; [http://www.ihtsdo.org/snomed-ct/].

27. Fan JW, Friedman C. Word sense disambiguation via semantic type classification. *AMIA Annu Symp Proc.* 2008:177-181.

28. Leroy G, Rindflesch TC. Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a naive Bayes classifier. *Stud Health Technol Inform.* 2004;107(Pt 1):381-385.

29. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in Health Technology & Informatics.* 2001;84(Pt 1):216-220.

30. IHTSDO. SNOMED Clinical Terms® Technical Implementation Guide. 2009.

31. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. *Journal of the American Medical Informatics Association : JAMIA.* 2005;12(4):486-494.

32. Browne AC, McCray AT, Srinivasan S. *The Specialist Lexicon*: National Library of Medicine;1993. NLM-LHC-93-01.

33. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* 2001:17-21.

34. Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proc Annu Symp Comput Appl Med Care.* 1994:240-244.

35. Gu HY, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: Modeling issues and advantages. *J Am Med Inform Assn.* Jan-Feb 2000;7(1):66-80.

36. Zong C, Halper M, Geller J, Peri Y. A structural partition of the Unified Medical Language System's semantic network. Paper presented at: Information Technology Applications in Biomedicine, 2000. Proceedings. 2000 IEEE EMBS International Conference on; 2000, 2000.

37. Melton GB, Moon S, McInnes BT, Pakhomov S. Automated Identification of Synonyms in Biomedical Acronym Sense Inventories. *The Louhi 2010: Workshop on Text and Datamining of Health Documents*. Los Angeles, CA2010:46-52.

38. Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J Biomed Inform.* Aug 2001;34(4):249-261.

39. McInnes BT, Pedersen T, Carlis J. Using UMLS Concept Unique Identifiers (CUIs) for word sense disambiguation in the biomedical domain. *AMIA Annu Symp Proc.* 2007:533-537.

40. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*. Toronto, Ontario, Canada: ACM; 1986:24-26.

41. Fellbaum C. *WordNet : an electronic lexical database*. Cambridge, Mass: MIT Press; 1998.

42. OpenNLP. [http://opennlp.apache.org/].

43. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Edmonton, Canada: Association for Computational Linguistics; 2003:173-180.

44. Mohammad S, Pedersen T. Combining lexical and syntactic features for supervised word sense disambiguation. Paper presented at: Proc of the CoNLL2004.

45. Long WJ. Parsing free text nursing notes. *AMIA Annu Symp Proc.* 2003:917.

46. Coden AS, GK; Buntrock, JD; Sominsky, IL; Ogren, PV; Chute, CG and de Groen, PC. Text Analysis Integration into a Medical Information Retrieval System: Challenges Related to Word Sense Disambiguation. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems.* Vol 129. Amsterdam2007:2218-2219.

47. Manning CD, Schütze H. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press; 1999.

48. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Assoc.* Jul-Aug 2004;11(4):320-331.

49. Ng HT, Lee HB. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. *Proceedings of the 34th annual meeting on*

Association for Computational Linguistics. Santa Cruz, California: Association for Computational Linguistics; 1996:40-47.

50. Denny JC, Spickard A, 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc.* Nov-Dec 2009;16(6):806-815.

51. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC Bioinformatics.* 2006;7:334.

52. Moon S, Pakhomov S, Liu N, Ryan J, Melton G. A comprehensive sense inventory for clinical abbreviations and acronyms using biomedical, biomedical literature, and medical dictionary resources2012. Located at: J Am Med Inform Assoc (submitted).

53. Moon S, Pakhomov S, Ryan J, Melton G. Automated Non-Alphanumeric Symbol Resolution in Clinical Texts. *AMIA Annu Symp Proc.* 2011:979-986.

54. Hunt DR, Verzier N, Abend SL, et al. *Fundamentals of Medicare Patient Safety Surveillance: Intent, Relevance, and Transparency*2005.

55. Wilkins LW. *Stedman's Medical Abbreviations, Acronyms & Symbols*. 4 ed2008.

56. NIH. Unified Medical Language System. 2010.

57. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. *Bull Med Libr Assoc.* Apr 1993;81(2):184-194.

58. McInnes BT, Pedersen T, Pakhomov SV. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *AMIA Annu Symp Proc.* 2009;2009:431-435.

59. Gale WA, Church KW, Yarowsky D. One sense per discourse. *Proceedings of the workshop on Speech and Natural Language*. Harriman, New York: Association for Computational Linguistics; 1992:233-237.

60. Leroy G, Rindflesch TC. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *Int J Med Inform.* Aug 2005;74(7-8):573-585.

61. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics (Oxford, England).* 2004;20(14):2320-2321.

62. Gale WA, Church KW, Yarowsky D. A Method for Disambiguating Word Senses in a Large Corpus. *Comput Humanities.* Dec 1992;26(5-6):415-439.

63. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care.* 1994:235-239.

64. Watson R. Part-of-speech Tagging Models for Parsing. Paper presented at: Pro of the 9th Annual Computational Linguistics community in the UK Colloquium.2006; Open University, Milton Keynes.

65. Yoshida K, Tsuruoka Y, Miyao Y, Tsujii Ji. Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers. *Proceedings of the 20th international joint conference on Artifical intelligence*. Hyderabad, India: Morgan Kaufmann Publishers Inc.; 2007:1783-1788.

66. Dell'orletta F. Ensemble system for part-of-speech tagging. Paper presented at: Proc of the 11th Conference of the Italian Association for Artificial Intelligence2009; Reggio Emilia, Italy.
67. Xu H, Fan J-W, Hripcsak G, Mendonca EA, Markatou M, Friedman C. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics.* April 15, 2007 2007;23(8):1015-1022.
68. Hwang FL, Yu MS, Wu MJ. The improving techniques for disambiguating non-alphabet sense categories. Paper presented at: Proc of Research on Computational Linguistics Conference XIII2000.
69. Yu MS, Hwang FL. Disambiguating the senses of non-text symbols for Mandarin TTS systems with a three-layer classifier. *Speech Communication.* 2003;39(3-4):191-229.
70. Jurafsky D, Martin JH. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall; 2000.
71. Willis MC. *Medical Terminology: The Language of Health Care*. 2 ed2006.
72. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications. Paper presented at: Proc of the 40th Anniversary Meeting of the Association for Computational Linguistics2002.
73. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 2009;11(1):10-18.

APPENDICES

Appendices generated from Chapter 2 are available at http://purl.umn.edu/137703 (website)

- The sense inventory of clinical abbreviations and acronyms (two versions)
- De-identified sentence datasets
- README (two versions)

Appendices generated from Chapter 4 are available at http://purl.umn.edu/137704 (website)

- De-identified sentence datasets
- README