IMPROVING INFORMATION EXTRACTION FROM CLINICAL NOTES WITH MULTIPLE DOMAIN MODELS AND CLUSTERING-BASED INSTANCE SELECTION

by

Youngjun Kim

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Computing

The University of Utah

December 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Youngjun Kim**

has been approved by the following supervisory committee members:

Ellen Riloff ,	Chair	**10/20/2017** Date Approved
Robert Kessler ,	Member	**08/30/2017** Date Approved
Suresh Venkatasubramanian ,	Member	**08/30/2017** Date Approved
John Hurdle ,	Member	**08/30/2017** Date Approved
Stéphane Meystre ,	Member	**09/20/2017** Date Approved

and by **Ross Whitaker** , Chair/Dean of

the Department/College/School of **Computing**

and by **David B. Kieda** , Dean of The Graduate School.

ABSTRACT

Extracting information from electronic health records is a crucial task to acquire empirical evidence relevant to patient care. In this dissertation research, I aim to improve two clinical information extraction tasks: medical concept extraction and relation classification.

First, my research investigates methods for creating effective concept extractors for specialty clinical notes. I present three new specialty area datasets consisting of Cardiology, Neurology, and Orthopedics clinical notes manually annotated with medical concepts. I analyze the medical concepts in each dataset and compare them with the widely used i2b2 2010 corpus. Then, I create several types of concept extraction models and examine the effects of training supervised learners with specialty area data versus i2b2 data. I find substantial differences in performance across the datasets, and obtain the best results for all three specialty areas by training with both i2b2 and specialty data. I also explore strategies to improve concept extraction on specialty notes with ensemble methods. I compare two types of ensemble methods (voting and stacking) and a domain adaptation model, and show that a stacked learning ensemble of classifiers trained with i2b2 and specialty data yields the best performance.

Next, my research aims to improve relation classification using weakly supervised learning. Due to limited labeled data and extremely unbalanced class distributions, medical relation classification systems struggle to achieve good performance on less common relation types, which capture valuable information that is important to identify. I present two clustering-based instance selection methods that acquire a diverse and balanced set of additional training instances from unlabeled data. The first method selects one representative instance from each cluster containing only unlabeled data. The second method selects a counterpart for each training instance using clusters containing both labeled and unlabeled data. These new instance selection methods for weakly supervised learning achieve substantial performance gains for the minority relation classes compared to supervised learning, while yielding comparable performance on the majority relation classes.

For my wife, Jaeshin.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Ellen Riloff, for her discipline throughout my graduate studies. I wish to express my deepest gratitude to her for broadening my knowledge of natural language processing, sparking my research interest in information extraction, and teaching me how to see and understand the big picture.

I am grateful for the guidance from Stéphane M. Meystre, whose expertise and knowledge in clinical NLP helped me achieve my goal. I am thankful to John F. Hurdle for his insightful perspectives and support of my efforts. I also greatly appreciate my other committee members, Bob Kessler and Suresh Venkatasubramanian, for providing productive feedback and helpful suggestions.

My heartfelt thanks to Vivek Srikumar and Hal Daumé III for their tremendously helpful advice and opening my eyes to new areas of NLP and machine learning. I am thankful to Jennifer H. Garvin for her kind support that has allowed me to keep my research work. Special thanks to Brian Roark for introducing me to the NLP field.

I appreciate Jennifer Thorne, RN and Jenifer Williams, RN for their annotation work. I sincerely thank Sean Igo and Olga Patterson for their help in data collection and processing. I would like to thank the valuable comments of all NLP lab members: Nathan Gilbert, Ruihong Huang, Lalindra De Silva, Ashequl Qadir, Haibo Ding, Jie Cao, Maks Cegielski-Johnson, Annie Cherkaev, Tianyu Jiang, Tao Li, Xingyuan Pan, Qinyun Song, Yichu Zhou, and Yuan Zhuang. I truly enjoyed our weekly NLP seminar. I wish to acknowledge the help provided by the ADAHF project team members, Julia Heavirland and Natalie Kelly. I truly appreciate Hua Xu, Berry de Bruijn, Colin Cherry, XiaodanZhu, and Bryan Rink for the prompt and kind responses to my inquiry. I would also like to thank my graduate advisors at the University of Utah School of Computing, Leslie LeFevre, Ann Carlstrom, and Karen Feinauer for their helpful advice.

This dissertation could not have been possible without the support and encouragement of my family. Words can't describe how thankful I am to my wife, Jaeshin Kwon, and my

children, Juyoung Kim and Giyoung Kim.

This research was supported in part by the National Science Foundation under grant IIS-1018314 and the National Library of Medicine under grant R01-LM010981. The de-identified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the NLP Shared Tasks organized by Dr. Özlem Uzuner and colleagues. I thank the i2b2/VA challenge organizers for their efforts, and gratefully acknowledge the support and resources of the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 Translational Use Case Projects; Utah CDC Center of Excellence in Public Health Informatics (Grant 1 P01HK000069-01), the National Science Foundation under grant IIS-1018314, and the University of Utah Department of Biomedical Informatics.

CHAPTER 1

INTRODUCTION

Natural language processing (NLP) has played an important role in automated electronic health records (EHR) mining. It has the potential to facilitate the analysis of massive EHRs and the acquisition of relevant patient information for improved data quality and decision making. The Centers for Disease Control (CDC) and Prevention survey in 2013 [109] reported that 78% of office-based physicians used some type of EHR system in the United States. Physician adoption of basic EHR systems was just 18% in 2001 so it has significantly increased. As manual curation of these EHRs is extremely expensive and time-consuming, a need for NLP technology is still growing in the medical community. Kaufman et al. [124] showed that incorporating NLP for EHR documentation was more effective than standard keyboard-and-mouse data entry in terms of documentation time, documentation quality, and usability. An evidence-based decision support tool utilizing EHRs, IBM Watson for Oncology (WFO) agreed 79% of the time with its human counterparts, Manipal multidisciplinary tumor board (MMDT), in diagnosing nonmetastatic disease [234].

EHRs contain important medical information related to patient care management. Health care professionals enter a patient's medical history and information about their care at a health care provider. A patient's diseases, symptoms, treatments, and test results are encoded in these notes in an unstructured manner. Meystre et al. [168] explained that some issues complicate the application of NLP on EHRs, such as ungrammatical phrases, prevailing usage of abbreviations and acronyms, and misspelled words. In this dissertation, the terms EHR, EMR (electronic medical record), and clinical note are used synonymously.

EHRs come in a variety of note types and are entered by health care professionals from varying backgrounds. In this research, I put the EHRs into two categories: *broad medical texts* and *specialty notes*. Broad medical texts, such as discharge summaries and progress notes, describe a patient's overall care and their content can cover a diverse set of topics

cutting across many areas of medicine. Broad medical texts have the advantage of being relatively well formatted, and they typically follow general documentation standards. The Joint Commission on the *Accreditation of Healthcare Organizations* (Hospital Accreditation Standards, IM.6.10, Elements of Performance 7) [119] recommends that a discharge summary should include the following information:

- The reason for hospitalization
- Significant findings
- Procedures performed, care, treatment, and services provided
- The patient's condition at discharge
- Information provided to the patient and family, as appropriate

Specialty notes are authored by medical specialists or medical-related occupations in the specialty divisions. In contrast to broad medical texts, specialty notes conform to varying documentation standards, with little overlap between specialties. They contain information emphasizing specific medical problems, specialized laboratory results, and clinical procedures. Specialty areas can be defined in several ways. They are often classified by organ system, patient group, or medical procedures. In the AMA (American Medical Association) physician specialty group and codes [7], 27 specialty areas are grouped and subspecialty areas are also listed for each specialty group. The contents of EHRs may vary across specialties. Physicians with a particular specialty need different information for each patient. The EHRs generated from Orthopedics are more likely to focus on the treatments for joint and muscle injuries, while the Cardiology notes would contain the more detailed descriptions of heart structure and function. Some contents might be missing from broad medical texts but they are recorded in specialty notes. Kripalani et al. [137] reported that 65% of test results pending at discharge were missing from discharge summaries in their observational studies. The restricted interaction between specialties may cause the variation of clinical language across specialty domains. Patterson and Hurdle [196] and Friedman et al. [90] demonstrated that clinicians in different clinical domains use specific sublanguages.

NLP techniques combining linguistic, statistical and heuristic methods have been applied to process unstructured texts. As a subfield of clinical NLP, information extraction (IE) has

been leveraged to extract structured information such as medical concepts from unstructured EHRs. Medical concept extraction (MCE) typically consists of two main steps: detection of the phrases that refer to medical entities, and classification of the semantic category for each detected medical entity. Leaman et al. [143] addressed that rich terminology describing medical concepts make MCE in the clinical domain challenging. Medical domain knowledge and sophisticated information extraction methods are often intergrated to achieve high levels of performance.

Labeled data need to be employed to train supervised machine learning algorithms for the MCE task. However, accessing clinical text corpora is more restricted than other biomedical text sources for reasons of confidentiality and de-identification requirements. In addition, most publicly available corpora of EHRs consist of broad medical texts. Annotated text collections representing specialty notes are less attainable. The reason for less availability of annotated specialty notes would be related to security or confidentiality. Grinspan et al. [95] addressed that confidentiality, workflow, and different needs for digitized information cause the variability of EHR use among specialties and the restricted exchange of clinical data across specialty areas. For the EHRs generated from certain specialties, additional protections are required when the notes contain particularly sensitive information of patient confidentiality. For example, psychiatry notes contain mental disorders related to patient privacy that needs to be protected more than other types of information. Therefore, the data annotated for the exclusive purpose of the intended users tends to be less sharable with external institutions. Chapman et al. [35] recognized six barriers to NLP development in the clinical domain and I believe that they especially apply to specialty areas. The barriers addressed by Chapman et al. [35] are as follows:

- Lack of access to shared data
- Lack of annotated datasets for training and benchmarking
- Insufficient common conventions and standards for annotations
- The formidability of reproducibility
- Limited collaboration
- Lack of user-centered development and scalability

My research goal aims to provide more flexible and robust integration of MCE models by ensemble-based approach on specialty notes with the limited amount of labeled data. I exploit stacked generalization [276], which is a meta-learning ensemble-based machine learning method, for the specialty areas. Stacked generalization uses the outputs of multiple learners to take advantage of their complementary strengths and diversities. I create multiple components learned from broad medical texts and specialty notes using a variety of IE techniques. Then, I combine broad medical components and specialty area components in a single ensemble to improve MCE on targeted specialty areas.

Medical concept extraction is a fundamental problem that can serve as the stepping stone for high-level tasks, such as inferring a medical concept's assertion, which is contextual information related to polarity, temporality, and relevance to the patient. The assertion information can play an important role in medical relation classification, another example necessitating MCE. For this research, I create a supervised assertion classifier to identify the assertion type of medical problems. For weakly supervised learning preparation of the medical relation classification task, the assertion classification model is applied to classify the assertion of each medical problem concept extracted from unlabeled data.

Medical relation classification (MRC) is one of the main tasks in this dissertation research. MRC involves involves recognizing different types of relationships between pairs of medical concepts. For instance, a relation can be extracted between a lab test and the test outcome it revealed. Identifying relations between concepts is essential to provide accurate and complete information about the concepts.

Classifying relations between pairs of medical concepts in clinical texts is a crucial task to acquire empirical evidence relevant to patient care. For example, extracting relations between mentions of a medication and mentions of allergy symptoms enables differentiation between situations when a medication causes the symptoms and situations when a medication is prescribed to alleviate symptoms.

When the amount of labeled data is small, achieving good performance on less common "minority" relation types is challenging. But infrequent relations can capture valuable information that is important to identify. For example, when a treatment is generally considered safe, it may result in side effects. Although a side effect may be rare for the specific treatment, recognizing it and providing a proper treatment is important to prevent

the unwanted encounters. My second research goal aims to improve medical relation classification in EHRs with an emphasis on minority classes by exploiting large amounts of unlabeled clinical texts, which are readily available in abundant quantity. I present two instance selection methods for more accurate classification of minority classes in the MRC task. These methods selectively choose unlabeled instances for self-training in an iterative weakly supervised learning framework. Both methods apply a clustering algorithm to group instances into clusters based on similarity measures. In the following sections, I give an overview of my research contributions for MCE and MRC.

1.1 Medical Concept Extraction for Specialty Notes

MCE is a challenging problem of growing interest to both the NLP and medical informatics communities. My research starts with the MCE task defined for the 2010 i2b2 Challenge [263]. This task involves extracting three types of medical concepts: problems (e.g., diseases and symptoms), treatments (e.g., medications and procedures), and tests.

The dominant research of MCE in the clinical domain has primarily been focused on broad medical texts. Most publicly available corpora of clinical medical notes consist of broad medical texts (e.g., i2b2 Challenge Shared Tasks [241, 259–263] and ShARe/CLEF eHealth Shared Tasks [125, 242]). I use the 2010 i2b2 corpus as broad medical texts. This corpus consists of discharge summaries and progress notes from various divisions in three different medical institutions.

There has been relatively little research on MCE for more specialized clinical texts. Studies focused on radiology and pathology reports are an important exception, but I would argue that they also cover a broad set of clinical conditions. Considering the current situation, creating corpora or text collections representing more diverse specialty areas would be valuable. A contribution of this research is that I created a new annotated data set of specialty notes from the BLUlab corpus. The BLUlab corpus is a large corpus of de-identified EHRs drawn from multiple clinical settings at the University of Pittsburgh Medical Center. I used the specialty area categorizations classified by Doing-Harris et al. [72]. They divided the BLUlab corpus into nine specialty groups for their specialty sublanguages research across institutions. For my research, I extracted three specialty areas from the corpus: cardiology, neurology, and orthopedics. To keep compatibility with the 2010 i2b2

corpus, two independent medical experts annotated the medical concepts using the 2010 i2b2 Challenge guidelines [3].

I investigate methods for creating MCE systems that will perform well on specialty area notes including ensemble learning for the MCE task with both broad medical texts and specialty area notes. The IE models and ensemble based methods are briefly described in the following section.

1.1.1 Information Extraction Models

I explore several classification methods, including rule-based, knowledge-based, and multiple machine learning models. Each method is capable of automated recognition of medical concepts without manual medical expertise. Four types of IE models that use a diverse set of extraction techniques are developed.

- MetaMap: I use a well-known knowledge-based system, MetaMap [10], that assigns UMLS Metathesaurus semantic concepts to phrases. The UMLS semantic types covering the types of medical concepts are effectively identified without any manual effort.

- Rules: I create a set of probability-associated rules involving word or phrase frequency aligned to medical concepts. These rules are compiled by harvesting information from the annotated training data. Domain expertise for the development of the rules is not required.

- Contextual Classifier: I create a supervised learning classifier with contextual features. I train a support vector machine (SVM) classifier with a linear kernel for multiclass classification.

- Sequential Classifier: I train sequential taggers using linear chain conditional random fields (CRF) supervised learning models. In contrast to the contextual classifier mentioned above, the CRF classifiers use a structured learning algorithm that explicitly models transition probabilities from one word to the next.

I examine these IE models and evaluate their performance on both broad medical texts and specialty notes. I investigate how well MCE models perform on specialty notes when

trained on a broad medical corpus and then when trained on the same type of specialty data.

1.1.2 Stacked Generalization for Medical Concept Extraction

For certain NLP tasks that already reach mature performance by a single classifier, ensemble methods have been used to further improve performance. These approaches are capable of producing more accurate predictions by regulating less accurate outputs from individual classifiers. They have been applied to overcome the shortcoming of an individual IE method dependent on a specific algorithm.

I explore two types of ensemble architectures that use the medical concept extraction methods described above as components of the ensemble. I created a voting ensemble, as a simple but often effective ensemble method, and a stacked generalization ensemble.

My stacked ensemble trains a meta-classifier with features derived from the predictions and confidence scores of a set of diverse component classifiers. Figure 1.1 shows the architecture of the stacked generalization ensemble. This ensemble architecture can be beneficial in two ways: (1) it can exploit multiple models that use different extraction techniques, and (2) it can exploit multiple models trained with different types of data.

I combine broad medical data and specialty data to outperform models trained on either type of data alone, when the amount of specialty data is limited. The stacked ensemble provides effective integration of any set of MCE models because it automatically controls the influence of each MCE model.

1.2 Exploiting Unlabeled Data for Relation Classification

Given a pair of medical concepts found in a sentence, a relation classification system must determine the type of relation that exists between the two concepts. My research focuses on the relation classification task introduced in 2010 for the i2b2 Challenge Shared Tasks [263]. This task involves recognizing eight types of relations between pairs of three types of medical concepts: problems, treatments, and tests. Note that this task aims to classify relations of given reference standard concepts.

The most successful methods used for relation classification include various supervised machine learning algorithms [263]. Extremely skewed class distributions pose substantial

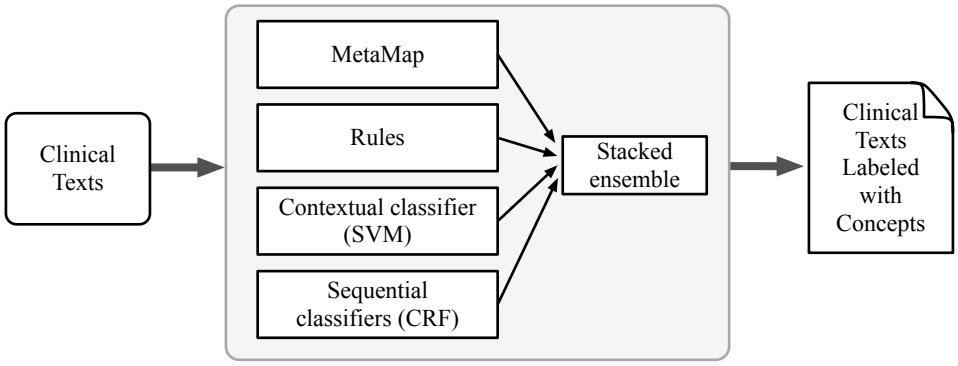


Figure 1.1: Architecture for a Stacked Generalization.

challenges for supervised machine learning because only a small number of labeled examples are available for training. As a result, machine learning classifiers can achieve high accuracy for the dominant classes but often perform poorly with the minority classes. Manually annotating more data is not a viable solution because of the high cost of manual annotation by medical experts. Also, because the minority classes are relatively rare, each batch of new annotations would provide only a relatively small number of new examples. There is a substantial cost for low reward.

As this time-consuming manual annotation effort poses a limited benefit, weakly supervised learning has been pursued to extend the amount of training data more efficiently. To take advantage of the large amounts of unlabeled clinical notes that are available, I explore an iterative weakly supervised learning framework. My research explores the idea of grouping both labeled and unlabeled instances together into clusters and using similarity measures to obtain new training instances by cluster analysis. I use the clustering-based instance selection methods to acquire a diverse and balanced set of additional training instances from unlabeled data. Figure 1.2 shows the process for a learning mechanism for medical relation classification exploiting unlabeled data. First, the relation classifier is trained only with labeled data. Second, the classifier is applied to the unlabeled data so that each unlabeled instance receives a predicted label. Third, new instances from unlabeled data are obtained through my new instance selection methods. Finally, the classifier is retrained with them.

1.3 Claims and Contributions

There are two claims that are to be made for this dissertation:

1. *Ensemble methods with a combination of models, some trained on broad medical texts and others trained on specialty area texts, can improve medical concept extraction on specialty notes.*

   Models trained with a combination of broad medical data and specialty data perform better than models trained on either type of data alone, when the amount of specialty data is limited. I present a way to combine multiple models that use different extraction techniques, and multiple models trained with different types of data. For this research, I create MCE models that use a diverse set of extraction techniques. Compared to MCE models trained on a broad medical corpus or trained on the same type of specialty

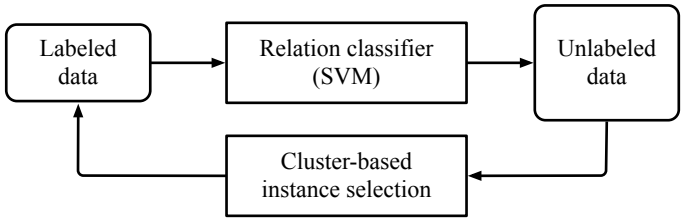


Figure 1.2: Learning Mechanism for Relation Classification.

data, an ensemble containing MCE models trained on broad medical texts as well as MCE models trained on specialty area texts achieves consistently better performance on specialty area notes.

A stacked learning ensemble with mixed domain models can achieve good performance with a favorable recall/precision trade-off. Compared to the model trained on the union of broad medical data and specialty data, a stacked learning ensemble consisting of a diverse set of MCE models trained from both broad medical data and specialty data can yield more precise extraction of medical concepts in specialty areas.

My stacked learning ensemble also offers the advantage of being able to easily incorporate any set of individual concept extraction components. The stacked learning ensemble, in which the meta-classifier automatically learns the beneficial effect of a new component, can offer more extensible and efficient integration of MCE models.

2. *Clustering-based instance selection from unlabeled data can improve performance on minority classes in medical relation classification.*

   This research shows that clustering-based instance selection from unlabeled text data outperforms supervised classification and traditional self-training on minority classes for relation type classification between medical concepts. I present two instance selection methods in weakly supervised learning specifically aimed at improving performance on minority classes. These two methods are based on clustering unlabeled data and can create a diverse and representative set of new instances from the unlabeled data.

   The first instance selection method, called *Unlabeled Data Prototypes (UDP) Selection*, selects instances from clusters containing only unlabeled data. The most representative instance from each cluster is selected as additional training data. The second method, called *Labeled Data Counterparts (LDC) Selection*, selects instances from the clusters containing both labeled and unlabeled instances. For each labeled instance, this method identifies its closest counterpart by selecting the unlabeled instance in the cluster that is most similar to it.

   This research also demonstrates that these new instance selection methods maintain good performance on the majority classes. These methods produce improvements on the majority classes with fewer selected instances than self-training. They can offer a

robust solution for classification problems when the data has a highly skewed class distribution, but large quantities of unannotated text data are available.

1.4 Guide to This Dissertation

The rest of this dissertation is structured as follows:

- Chapter 2 is devoted to discussing related work to this dissertation research. It describes how concepts have been extracted in newswire articles, biomedical literature, and clinical notes; how the assertion information of a medical concept has been determined; and how relations between a pair of concepts have been classified in these NLP domains. Then, this chapter explores how the components of information extraction methods have been incorporated and how unlabeled data has been exploited.

- Chapter 3 describes medical concept extraction from specialty area clinical notes. This chapter presents new text corpora created from three specialty areas and analyzes their difference in content from each other and i2b2 medical notes. It discusses a variety of information extraction models and evaluates their performance on all of these data sets. This chapter also investigates how ensemble-based methods that combine multiple MCE models perform on specialty notes.

- Chapter 4 describes details of assertion classification and the features used for this task. Then, it presents the classification results with the full set of features and also investigates the improvements resulting from the addition of each type of features.

- Chapter 5 focuses on medical relation classification and discusses the challenges of this task. It describes data preparation for weakly supervised learning by identifying the medical concepts in the unlabeled data and classifying the assertion of each medical problem concept. This chapter introduces supervised classification models, self-training with confidence-based instance selection, and two clustering-based instance selection methods. Then, it presents the experimental results and compares the differences between these selection methods. This chapter demonstrates that clustering-based instance selection methods from unlabeled text data improve performance on minority classes for relation type classification.

- Chapter 6 summarizes this dissertation. It discusses the research contributions and outlines directions for future research.

CHAPTER 2

RELATED WORK

In the area of clinical NLP, a large number of studies have emerged to address the challenging task of *information extraction* (IE). *Medical concept extraction* (MCE) and *medical relation classification* (MRC) tasks are analogous to *entity recognition* and *relation extraction* tasks in the general language domain respectively. Both tasks are often considered as subtasks of information extraction, which is defined as "*the process of scanning text for information relevant to some interest, including extracting entities, relations, and, most challenging, events—or who did what to whom when and where*" [107].

In this chapter, I present related work to this dissertation research. I describe (1) how concepts have been extracted in newswire articles, biomedical literature, and clinical notes (Section 2.1), (2) how the assertion information of a medical concept has been determined (Section 2.2), and (3) how relations between a pair of concepts have been classified in these NLP domains (Section 2.3). Then, I explore how the components of IE methods have been incorporated and unlabeled data has been exploited with regard to these target tasks (Section 2.4, Section 2.5).

2.1 Recognizing Concepts

I examine the question of how the semantic types of concepts have been defined for different target domains. I briefly introduce the history of the named entity recognition (NER) task and present influential research.

2.1.1 Named Entity Recognition From General Text

Once originated from MUC-6 (the sixth in a series of Message Understanding Conferences) [97], NER has been a critical issue in IE along with relation extraction (RE) and coreference resolution. The goal of NER is to extract and classify proper named or specialized entities into predefined categories such as *Person*, *Organization*, *Location*, time expressions (*Date*, *Time*),

and number expressions (*Money*, *Percent*) [97]. Collections of newswire articles written in English or other languages have been presented for several shared tasks [41, 71, 253, 254] on NER. In the CoNLL conferences [253, 254], four types of entities were defined: *Person*, *Organization*, *Location*, and *Miscellaneous* names. The Automatic Content Extraction (ACE) Program [71] called for recognizing seven types of entities: *Person*, *Organization*, *Location*, *Facility*, *Weapon*, *Vehicle*, and *Geo-Political Entity*.

Since the beginning of the new millennium, there have been outstanding achievements in the extraction of named entities (NEs) and various machine-learning algorithms have been applied. Mikheev et al. extracted NEs from MUC-7 [170] articles using the combination of rule-based grammars and statistical models. They conducted a series of experiments with different settings of gazetteers, dictionary lists of people, organizations, and other NEs. They observed that their NER system could yield satisfactory performance on Person and Organization entities, even without gazetteers, because of the contextual clues available in texts. For example, "***XX**, the CEO of **YY***" as evidence ***XX*** is a *person* and ***YY*** is an *organization*.

Recent NER systems have been framed as a sequential token-based labeling problem with various encoding schemes to assign a class label to each word in a sequence. Borthwick et al. [24] allowed maximum entropy models to be trained with the combined features from training data and the outputs of other NER systems [96, 138, 152]. They reported that incorporating the outputs of external systems outperformed the best individual classifier. Lafferty et al. [141] presented the conditional random fields (CRF) learning method to build a statistical model for sequential labeling. They showed CRF models could overcome the independence assumption problems in generative models such as hidden Markov models (HMMs) [17, 204] and provide a more robust solution over maximum entropy Markov models (MEMMs) [160]. McCallum and Li [161] used CRFs to perform NER and their system achieved an 84.0% F_1 score on the CoNLL-2003 English test set [254].

Other machine-learning techniques also have been used for the NER task. Collins [50] describes the perceptron algorithm [218] for training tagging models, as an alternative to MEMMs and CRFs. Collins [51] also used the voted perceptron [88] to rerank the N-best sequences of labels outputted from an MEMM tagger and the reranking method showed a significant improvement over the MEMM tagger. Isozaki and Kazawa [113] and Asahara

and Matsumoto [11] used the support vector machines (SVM) [53, 267] algorithms to extract NEs from Japanese texts.

Some research relevant to this dissertation research can be found in semantic class induction (SCI) methods. The goal of SCI is to construct a lexicon (dictionary), lists of words with semantic class labels. The target entity types of SCI are less strict than NER and include nominal noun phrases. NER classifies instances in context but SCI produces stand-alone lexicons.

Bootstrapped learning, a form of *self-training*, has been widely applied to build semantic lexicons [102, 211, 214, 248, 275]. Starting with a set of seed words, bootstrapping labels matching entities in a corpus. It uses pattern contexts to identify new entities. This iterative procedure often causes semantic drift [59] that occurs when the extracted entities do not correctly represent the original semantic class. To reduce the semantic drift problem, McIntosh and Curran [166] proposed a filtering method based on the distributional similarity between extracted entities.

Recently, Huang and Riloff created semantic class taggers induced by bootstrapping that exploited a domain-specific corpus for veterinary medicine [112]. Then, Qadir and Riloff [202] proposed an ensemble method incorporating pattern-based bootstrapping, the semantic taggers of Huang and Riloff [112], and coreference-based lexicon construction. Their results showed that the ensemble-based approach acquired higher quality of lexicons and obtained better semantic tagging results than the individual methods.

In this section, I covered NER for the general text domain by describing the entity types of interest to be extracted, shared tasks focusing on the NER task, and the techniques proposed for the task. Also, I briefly reviewed research that focused on semantic class induction. A more comprehensive review of general text NER can be found in the review by Nadeau and Sekine [181]. In the following two subsections, I describe the studies targeting NER from biomedical literature and EHRs.

2.1.2 NER From Biomedical Literature

NER in biomedical natural language processing (BioNLP) has advanced from the general text NER by sharing the algorithms and features. In BioNLP, IE has aimed at extracting genetic information or bio-entities, such as *Gene*, *DNA*, *RNA*, and *Protein* from biomedical

literature. NER in BioNLP can be confronted with some difficulties. As pointed out by Tuason et al. [257] and Yeh et al. [284], (1) gene mentions are often made up of common words instead of proper nouns, (2) new genes are continually being discovered and known ones are renamed, (3) in many cases, they are named without following standard naming conventions, (4) genomic databases cannot cover the entire biological entities.

Several shared tasks [105, 130, 131, 208, 284] related to BioNLP have been organized and corpora with bio-entity annotations have been developed. The *BioCreAtIvE* task 1A [284] dealt with gene name detection. The target gene entities for this task includes *binding sites*, *motifs*, *domains*, *proteins*, *promoters*, and other information, but they are identified as "names" with no distinction between them. In the Bio-Entity Recognition Task at the *JNLPBA* shared task [130], five different types of bio-entities, *protein*, *DNA*, *RNA*, *cell line*, and *cell type*, were targeted for NER. Although genes and proteins are the most common entities annotated in biomedical corpora, other semantic types are also used for entity recognition. Nerves [185] analyzed the semantic annotations from 36 biomedical corpora and categorized the semantic types into the six following groups: *gene/protein*, *variant/mutation*, *drug/chemical*, *cell/anatomy*, *disease*, and *organisms*.

Similar to NER in the general text domain, diverse machine-learning algorithms have been applied for the BioNLP NER. To extract proteins or genes from the *BioCreAtIvE* corpus or the *JNLPBA* corpus, HMMs [291, 295], MEMMs [84, 85], CRFs [164, 227], or combinations with multiple models [235, 297] have been employed. Some NER systems originally trained for general texts were adapted for the BioNLP domain. Zhou and Su [296] proposed an HMM-based NER tagger for MUC-6 and MUC-7 and their tagger [297] was altered for *GENIA* V3.0 corpus [129]. For the *BioCreAtIvE* and *JNLPBA* evaluation, Finkel et al. [84, 85] customized the system developed for the CoNLL shared task [134] with domain specific features.

I reviewed the work aimed for NER on biomedical literature and the semantic types defined for this domain. I did not attempt a complete review of NER in BioNLP, such reviews are available in [48, 104, 110, 232]. In next subsection, I describe the research related to medical concept extraction from EHRs.

2.1.3 Medical Concept Extraction From EHRs

MCE is one of the fundamental tasks to transform free narrative text data in EHRs into structured information. Recent work in the clinical domain has demonstrated that NLP has the potential to process EHRs for concept detection even though there exists additional challenges due to limited access to shared data, the unstructured format of narrative texts, and heterogeneous contents across various medical areas and health care providers.

In early NLP research for clinical notes, most systems relied on rule-based approaches. Friedman et al. [89] created *MedLEE* which has been applied to chest radiology reports, discharge summaries, and operative reports to extract and encode medical information. *MedLEE* [89] uses a rule-based system that extracts medical concepts by performing a shallow syntactic analysis and using semantic lexicons. *SymText* was developed by Haug et al. [99, 100] and evolved into *MPlus* [44]. *MPlus* [44] was used to extract medical findings, diseases, and appliances from chest radiograph reports. Heinze et al. [103] presented *LifeCode* to extract demographic and clinical information from EHRs in emergency medicine and radiology specialties. *MetaMap* [10] was developed to recognize Metathesaurus concepts from biomedical texts by utilizing the *UMLS* (Unified Medical Language System) [22, 154]. The *UMLS* is a repository of health and biomedical vocabularies and standards developed by the US National Library of Medicine [22]. *MetaMap* has frequently been used for EHRs as well. Zou et al. [306] presented syntactic and semantic filters to remove the irrelevant concept candidates in the extraction of key UMLS concepts from EHRs.

Many applications have used open standard frameworks such as UIMA (Unstructured Information Management Architecture) [82, 83] and GATE (the General Architecture for Text Engineering) framework [57, 58], which have improved scalability and interoperability between different analytical components. Zeng et al. [286] built *HITEx* (Health Information Text Extraction), which is a GATE pipelined system with multiple preprocessing modules, to extract family history information, principal diagnosis, comorbidity and smoking status from clinical notes. For the classification of subjects with rheumatoid arthritis, Liao et al. [150] used *HITEx* to extract clinical information such as disease diagnoses, medications, laboratory data, and radiology findings of erosions. Savova et al. [224] built an open-source UIMA based IE application, *cTAKES* (the clinical Text Analysis and Knowledge Extraction System), consisting of a number of components trained for the clinical domain, such as

a drug NE recognizer, assertion classifier, constituency parser, and so forth. It has been used for more than 20 research projects related to clinical NLP tasks including document classification [151], drug side effect extraction [233], coreference resolution [173, 293], and clinical question answering [31, 191]. Readers interested in clinical NLP can be referred to the detailed reviews of earlier and recent studies by Meystre et al. [168] and Nadkarni et al. [182].

Most current IE systems in clinical NLP use statistical machine learning approaches that often achieve better performance than rule-based approaches that typically require manual effort. MCE has been the focus of several shared tasks, such as the *i2b2 Challenge Shared Tasks* [262, 263], the *ShARe/CLEF eHealth Shared Tasks* [242]. My research focuses on the MCE task that was introduced in 2010 for the *i2b2 Challenge Shared Tasks* [263]. These challenge tasks included: (1) the extraction of medical problems, tests, and treatments, (2) classification of assertions pertaining to medical problems, and (3) relations between medical problems, tests, and treatments. The best performance on the 2010 i2b2 concept extraction task (1) was achieved by de Bruijn et al. [64] with 83.6% recall, 86.9% precision, and 85.2% F_1 score. They integrated many features commonly used in NER tasks including syntactic, orthographic, lexical, and semantic information (from various medical knowledge databases). Jiang et al. [116] trained a sequence-tagging model that consisted of three components in a pipeline: concept taggers with local features and outputs from different knowledge databases, post-processing programs to determine the correct type of semantically ambiguous concepts, and a voting ensemble module to combine the results of different taggers. Their system achieved an 83.9% F_1 score. Subsequent research by Tang et al. [247] showed that clustering and distributional word representation features achieved a higher F_1 score of 85.8%.

As discussed in this subsection, some research focusing on specialty notes has been proposed but most corpora annotated for specialty areas have not been shared with others. Most publicly available corpora provided by some shared tasks are selected from broad medical texts. Consequently, relatively little research on MCE from specialized clinical texts has been conducted because of the lack of shared datasets consisting of specialty notes. Effective integration of MCE models trained on both broad medical texts and specialty notes is presented in this dissertation.

I reviewed which types of concepts have been extracted in each domain and what methods

have been implemented for concept or entity recognition. The next section gives a brief overview of research that has tackled the assertion classification problem.

2.2 Classifying a Concept's Assertion Information

Taking into account the contextual information of a medical concept is crucial to acquire the practical meaning of the concept entered in EHRs. When a patient has multiple medical symptoms and they are transcribed in an unstructured manner, the contextual information of each symptom might vary. Distinguishing negated medical concepts from positive concepts would be important to summarize the patient's health information and the negated concepts need to be treated with extra care. When some clinical diagnoses are written as part of decision support messages or reminders, they are not associated with the actual patient, but rather serve as a recommendation. For example, "*Please contact primary care provider if the patient has recurrent fevers*" does not indicate the patient' medical problem and the medical problem, "*fevers*," should not be assigned to the patient.

Inferring the contextual information can be investigated as a multiclass problem and this task has been recently explored in the medical domain. For local context recognition and analysis, several NLP systems have been developed that focused on the negation or other assertions of medical concepts. For negation classification, rule-based systems like *Negfinder* [180] and *NegEx* [33] have been introduced. They used regular expressions with trigger terms to determine whether a medical term was negated. *NegEx* [33] reached 77.8% recall and 84.5% precision for 1,235 medical problems in discharge summaries. In the *BioNLP-2009* [128] and *CoNLL-2010* [81, 268] Shared Tasks, detecting negations (and their scope) in natural language text was the focus. Kilicoglu and Bergler [126] compiled negation cues from the corpus and detected the negation using dependency-based heuristics. Morante et al. [179] implemented two stages of negation scope detection: sentence level classification and phrase level with memory-based learning. After recognizing the negation signals from each sentence, they sought the full scope of these negation signals in the sentence.

Chapman et al. [34] introduced the *ConText* algorithm, which extended the *NegEx* algorithm to detect four assertion categories: *absent*, *hypothetical*, *historical*, and *not associated with the patient*. Uzuner et al. [264] developed the Statistical Assertion Classifier (*StAC*) and showed that a machine learning approach for assertion classification could

achieve results competitive with their own implementation of the Extended NegEx algorithm (*ENegEx*). They used four assertion classes: *present*, *absent*, *uncertain in the patient*, or *not associated with the patient*.

Assertion classification was the focus of the 2010 i2b2 NLP challenge [263] and the defined task consisted of choosing the status of medical problems by assigning one of six categories: *present*, *absent*, *hypothetical*, *possible*, *conditional*, or *not associated with the patient*. The best performing system [64] reached a microaveraged F_1 score of 93.6%. Their breakdown of F_1 scores on the individual classes was: *present* 95.9%, *absent* 94.2%, *possible* 64.3%, *conditional* 26.3%, *hypothetical* 88.4%, and *not associated with the patient* 82.4%. Several other studies then used this challenge data and showed that machine learning approaches [16, 46, 133] performed better than handcrafted rule-based systems. Kim et al. [133] used a variety of linguistic features, including lexical, syntactic, lexico-syntactic, and contextual features for assertion classification. They developed some features to improve the performance of minority classes. For example, as medical problems associated with allergies are annotated as *conditional* [2], five allergy-related section headers (i.e., "*Allergies*," "*Allergies and Medicine Reactions*," "*Allergies/Sensitivities*," "*Allergy*," and "*Medication Allergies*") were defined.

Some studies have been also carried out on similar specialized classification tasks, such as tumor status [40], lung cancer stages [188], and medication prescription status classification [167]. In the next section, I briefly introduce the relation extraction task in general text and the BioNLP domain, and describe how the medical relation classification work in this dissertation relates to prior research.

2.3 Classifying a Relation Between a Pair of Concepts

While concept recognition undertakes the detection of an individual entity of interest, and assertion classification infers the contextual attribute of a concept, relation classification has to deal with a pair of concepts. Thus, the relation classification task can be more challenging than those tasks described in the two previous sections (Section 2.1, Section 2.2), as it has to consider the pair of concepts and any information around them simultaneously. The relation classification is to determine whether a pair of concepts (or entities) are in a relation, and how they are semantically related when a relation exists between them. This task has been commonly addressed by two approaches in supervised learning: feature-based and

kernel-based. Using unlabeled data, weakly-supervised learning (including bootstrapping methods) has also been applied for this task.

In this section, I discuss the application of these methods in general NLP, bioNLP, and clinical NLP. I mainly focus on relations between exactly two concepts. (Note that the number of semantic classes does not necessarily need to be two.) For a more comprehensive review including higher-order relations [165], which extract complex relations between more than two concepts, the reader can refer to [13, 77, 178].

2.3.1 Relation Extraction in General Text

Similar to the NER task, relation extraction (RE) firstly commenced in general text NLP. This task was introduced as an NLP task for the *MUC-7* (Message Understanding Conference) [41, 158] and also added in Phase 2 of *ACE* (Automatic Content Extraction) [237]. In the MUC-7, three relation types were defined: *LOCATION_OF*, *EMPLOYEE_OF*, and *PRODUCT_OF*. The relations were extended for ACE and six relation types were defined in the ACE 2008 [12]: *Physical*, *Part-whole*, *Personal-Social*, *ORG-Affiliation*, *Agent-Artifact*, and *Gen-Affiliation*. Subtypes were also added for each major relation type. For example, the *ORG-Affiliation* relation has the following subtypes: *Employment*, *Ownership*, *Founder*, *Student-Alum*, *Sports-Affiliation*, *Investor-Shareholder*, and *Membership*.

Lin and Pantel [153] defined a *path* as a relation between two entities. They computed the statistics of paths derived from dependency trees in a corpus to find the similar paths (or "inference rules") of a given path. For example, for "*X is author of Y*," "*X writes Y*," and "*X co-authors Y*" were some of the paths correctly found. Miller et al. [172] used augmented parse trees to integrate syntax and semantic information of entities and relations. Their integrated model was jointly trained for RE and other NLP tasks to prevent the error propagation through the system pipeline.

To create relation classifiers from supervised training data, feature-based or kernel-based methods have been attempted with the datasets from these NLP shared tasks. In feature-based methods [98, 115, 121, 292], Kambhatla [121] constructed maximum entropy models employing lexical, syntactic and semantic features derived from the syntactic parse tree and the dependency tree. GuoDong et al. [98] also examined the combination of these kinds of features using SVM classifiers. They showed that chunking features regarding the phrase

heads and the phrase path were very effective for RE. Jiang and Zhai [115] conducted a thorough experimental evaluation with features used in previous work and new features in order to study the effectiveness of these features. They reported that adding complex features, such as complete subtrees of the syntactic parse tree, could hurt the performance when effective features are already included.

As an alternative to feature-based methods, Zelenko et al. [285] proposed kernel methods for RE. The kernel methods [30, 56, 285, 287, 288] focus on the construction of kernel (or similarity) function to compute the similarity between two relation instances. They used shallow parsing information to represent the relation instances and applied kernels that represent the similarity of two shallow parse trees. Zhang et al. [288] presented a composite kernel consisting of an entity kernel and a convolution parse tree kernel. They showed that the composite kernel is capable of using effective syntactic structure features. Qian et al. [203] pointed out that prior work using tree kernels [287, 298] may contain unnecessary information and miss useful context-sensitive information related to predicate-linked paths. They proposed a new method that exploits constituent dependencies to overcome these problems.

Some studies based on weakly supervised learning have been proposed for RE. Zhang [290] presented a bootstrapping algorithm using random feature projection. Multiple classifiers were trained with randomly selected features from labeled data and they voted to assign the labels of unlabeled data. Xu et al. [279] also used a bootstrapping algorithm to extract relations by applying the rules induced from linguistic patterns from labeled data. The patterns were derived from dependency trees containing seed examples. Chen et al. [38] proposed graph-based weakly supervised learning using a label propagation algorithm. Sun et al. [238] presented a weakly supervised learning method with large-scale word clustering. They augmented the features derived from the word clusters to compensate for the absence of lexical features in labeled data. Wang and Fan [271] collected training data using a clustering algorithm. To minimize the manual annotations, the most representative instance with the highest average similarity to other members of each cluster was selected for annotation.

Methods using *Distant Supervision* [54] have been popularly applied to RE. Distant supervision finds examples of the targeted relations from existing knowledge databases and uses them to generate the training instances. *Freebase* [23], an open large semantic

community-curated knowledge base, has been commonly used for RE. For example, using the fact that *The Incredibles* was directed by *Brad Bird* extracted from *Freebase* relations, the entity pair of "*The Incredibles*" and "*Brad Bird*" can be annotated as a *Directed_by* or *Director_of* relation when they co-occur in the same sentence. Mintz et al. [174] used Freebase to collect training instances from Wikipedia articles with distant supervision. Yao et al. [281] applied *selectional preferences* [209, 210] of relations to filter instances extracted from Wikipedia or New York Times articles. Instances that violated the selectional preferences were excluded to improve precision. More recent studies with distant supervision have been proposed: extending the knowledge resources [190], applying multi-instance multi-label learning [245], filtering out falsely labeled instances through pattens derived from the dependency parse [246], and combining with active learning [8].

There has been some research exploiting texts on the web to extract semantic relations. Etzioni et al. [78] developed an unsupervised IE system capable of scaling up relation extraction by collecting information from the web. Starting with a small set of generic extraction patterns, their system iteratively extracted entities and their relations and then used mutual-information statistics to validate them. This task was specified as Open Information Extraction (OIE) by Banko et al. [15]. Different than conventional RE that focused on predefined semantic relations, OIE attempts RE with nonrestricted (or all possible) types of relations from the web documents. Further studies using this domain-independent unsupervised method have been conducted: by processing Wikipedia articles [277], incorporating syntactic and lexical constraints to reduce uninformative and incoherent extractions [79], and extracting relations expressed by verbs and other syntactic entities as well [159].

I have discussed feature-based and kernel-based RE methods, weakly supervised approaches including Distant Supervision, and Open Information Extraction. In the next subsection, I review the research conducted for several RE subtasks in the biomedical domain.

2.3.2 Relation Extraction From Biomedical Literature

Along with entity recognition from biomedical literature, relation extraction has been an important topic in BioNLP as well. Biomedical literature has been used for both relation extraction and entity recognition because new medical entities and their relationships are

increasingly introduced in the biomedical literature but they are not often covered by knowledge-based databases. I survey the studies related to RE from biomedical literature in a rather different way than the previous subsection. While I discussed the RE in general text from the methodology perspective, the analysis of RE from biomedical literature would be focused on the specific tasks defined for BioNLP.

RE methods typically capture the relationships between pairs of biomedical entities as well. Several shared tasks focusing on RE including *protein-protein interaction* (*PPI*) [135, 136, 144], *gene interaction* [25, 183], and *drug-drug interaction* (*DDI*) [225, 226] have facilitated this research and more approaches have been proposed as post-shared task efforts.

Extraction of protein-protein interaction is a widely studied area in BioNLP. According to De Las Rivas and Fontanillo [65], PPI can be defined as a physical contact with molecular docking between proteins that occur in a cell or in a living organism. Blaschke et al. [18] used pre-specified protein lists and action verbs to identify the sentences containing PPI information from Medline abstracts. Pustejovsky et al. [201] created a parser to extract *inhibition*-relations, for example, "*the tail receptor peptide inhibits HGF-mediated invasive growth*". They applied anaphora resolution to extract more relations by comparing syntactic and semantic features between the anaphor and the candidate antecedents. Other research employing shallow or full parsing information includes [195, 249, 280]. Recently, Miwa et al. [175] combined different types of kernels based on various parse structures for PPI extraction. They also analyzed the compatibility of five PPI corpora and showed that the corpora could be adapted with a small effort to improve PPI extraction performance. Miyao et al. [176] presented a comparative evaluation of state-of-the-art parsers in PPI extraction. They obtained higher accuracy than the previous state-of-the-art results when they combined the deep parser by Miyao and Tsujii [177] and Charniak and Johnson's reranking parser [36].

Genic Interactions extraction was introduced in the *Learning Language in Logic challenge* (*LLL05*) [183]. The task is to extract protein/gene interactions from biology abstracts. More specifically, pairs of an agent and target with genic interactions, such as *action*, *binding and promoter*, and *regulon*, need to be extracted. Fundel et al. [92] created a rule-based system using dependency parse trees that provide non-local dependencies within sentences. Even with a small number of simple rules applied to these trees, they achieved competitive performance on the LLL challenge dataset. Kim et al. [132] found the shortest path between

two entities in the dependency relations to train multiple kernels. This shortest path allowed the kernels to be extended from flat linguistic features to structural information.

DDIExtraction-2011 [226] and *DDIExtraction-2013* [225] challenge tasks aim to extract drug-drug interactions from biomedical texts. Thomas et al. [250] obtained higher results than any other participants in DDIExtraction-2011 by implementing the majority voting of several kernel-based classifiers and a case-based reasoning (CBR) [5] system. The best performance on the DDIExtraction-2013 task was achieved by Chowdhury and Lavelli [43]. They used the scope of negation cues and semantic roles to filter less informative negative instances. For example, the sentence containing "*not recommended*" was removed before both training and testing stages.

Some studies have also been conducted for extraction of other types of relations. The relations of interest include relations between cancer drugs and genes [212], gene-disease relations [29, 45, 147], treatment-disease relations [29, 216], and protein-residue association [207]. The reader can find a more comprehensive survey of BioNLP relation extraction in the reviews of Zhou and He [294] and Tikk et al. [251]. While RE in BioNLP has primarily dealt with articles written by biomedical scholars or professionals, clinical NLP has focused on RE from EHRs written by healthcare practitioners and professionals. In the next subsection, I discuss the classification of relations between pairs of medical concepts occurring in EHRs.

2.3.3 Medical Relation Classification

The medical relation classification (MRC) task was defined as part of the Fourth i2b2/VA Shared Task Challenge [263] in 2010. Although some studies [9, 215, 299] were previously conducted with different data sets, the i2b2 2010 challenge data that is publicly available facilitated more research on the MRC task.

My research involves MRC for pairs of medical concepts, assuming that the terms corresponding to the two concepts have already been identified. The task is to identify how medical problems relate to treatments, tests, and other medical problems in clinical texts. Many sentences contain multiple pairs of concepts, so the challenge includes identifying which pairs are related, as well as identifying the specific type of relation.

Previous MRC work has presented microaveraged F_1 scores, which assess performance over all of the positive instances regardless of which class they belong to. However, microaveraging

obscures performance differences across the classes. For example, it is often possible for a system to achieve a high microaveraged F_1 score by performing well on the majority class but recognizing few, if any, instances of the minority classes. My research aims to shed light on the performance differences across relation classes, with the goal of comparing the ability of different methods to recognize the minority classes. So I will discuss both microaveraged and macroaveraged results in the rest of this subsection.

The MRC task has been typically cast as a feature-based supervised learning problem, where a classifier is trained with manually annotated data. Rink et al. [213] used supervised learning to produce the highest microaveraged F_1 score, 73.7%, for this relation extraction task. Their system utilized external resources including Wikipedia, WordNet, and the General Inquirer lexicon [236] as part of their feature set. To improve recall, they set much lower weights for the pairs of nonrelated concepts (i.e., negative examples) when training an SVM classifier. Their system reached macroaveraged metrics (not officially reported in Rink et al. [213] but calculated by taking the average of the reported recall and precision of the different subclasses) of 51.7% recall, 55.8% precision, and 53.7% F_1 score.

de Bruijn et al. [64] explored effective features also applicable to other clinical NLP tasks. In addition to supervised classification, they applied self-training on the provided unlabeled data. Their approach produced a 73.1% microaveraged F_1 score. Macroaveraged metrics for their submission reached 43.7% recall, 66.8% precision, and 51.2% F_1 score. I calculated these numbers based on the output of de Bruijn et al. [64]. Their subsequent research [302] using composite-kernel learning improved the accuracy of relation classification yielding a higher microaveraged F_1 score of 74.2%. As an effort to overcome the class imbalance problem, they used down-sampling of negative examples before training the models. D'Souza and Ng [74] presented an ensemble approach exploiting human-supplied knowledge to set up individual classifiers. Their weighted-voting system outperformed a single classifier using the full set of features exploited by different ensemble members. Their best-scoring ensemble system produced a 69.6% microaveraged F_1 score. Note that their result is not directly comparable with the works described above because of different training data sizes.

Even though several weakly supervised approaches have been proposed in general texts and BioNLP domains for relation extraction, no study exploiting large amounts of unlabeled clinical texts has been attempted for the MRC task. Applying self-training is an exception,

but it has only been implemented as one-shot self-training [64]. My research specifically aims to improve the classification of minority classes by clustering-based instance selection from unlabeled data. Benefiting from large amounts of unlabeled data with new instance selection methods based on similarity measures is a novel contribution of this dissertation.

So far I have described the research for concept extraction, classification of medical assertions, and relation extraction, as can be seen in the three previous sections (Section 2.1, Section 2.2, Section 2.3). I may now proceed to the discussion of combining the components of IE methods in ensemble based methods and exploiting unlabeled data in weakly supervised learning.

2.4 Combining Models

Ensemble methods that combine multiple classifiers have been widely used for many NLP tasks and generally yield better performance than individual classifiers. For the more effective ensembles, a diverse set of classifiers using different types of learning algorithms or data have been used [140, 198, 258]. Dietterich [69] addressed three reasons to construct good ensembles: to reduce the risk of using an inadequate one, provide a better approximation by divergent models, and possibly expand the space of representable functions by the combination of multiple models.

In this section, I describe research using ensemble methods that led to more accurate classification. Related work using voting ensemble and stacked generalization is covered in the next two subsections. Then, I briefly cover domain adaptation, which is also related to my research as domain adaptation often combines multiple models either from the source domain or the target domain.

2.4.1 Voting Ensembles

The voting ensemble has attracted NLP researchers because it can offer a convenient and often effective way to combine multiple predictive models without retraining a new model. It has been applied to several NLP tasks including Part-of-Speech tagging [27, 66, 265], phrase chunking [139, 221], word sense disambiguation [28, 197], relation classification [42, 74, 250], sentiment analysis [60, 269] and named entity recognition (NER), which is most closely related to this dissertation work.

In NER, Zhou et al. [295] used majority voting from multiple classifiers to achieve better performance than any single classifier. Finkel et al. [84] combined the outputs of forward and backward (reversing the order of the words in a sentence) sequence labeling, which improved recall. Li et al. [149] proposed multiple combinations with the forward and backward labeling of CRF and SVM models. They showed that the union model successfully improved performance. Similarly, Huang et al. [111] integrated the outputs of three models for gene mention recognition. They intersected the outputs of forward and backward labeling SVM models and then unioned with the outputs of one CRF (conditional random fields) model. Ekbal and Saha [76] pointed out that the weights of voting should vary among the different semantic types in each classifier in weighted voting. They introduced a new method of determining the weight of votes for all semantic types per classifier and their method was evaluated for Indian languages NER.

In clinical NLP, Doan et al. [70] showed that a voting ensemble of rule-based and machine learning systems obtained better performance than individual classifiers for medication detection. For the MCE task, Kang et al. [122] used majority voting between seven different systems for performance improvement. Their work is similar to my MCE voting ensemble although they had a different tiebreaker policy. When overlapped concepts have the same votes, they randomly select one concept while I pick the one with the highest confidence score.

2.4.2 Stacked Generalization

Stacked generalization (SG) is another ensemble-based method for combining multiple classifiers by training a meta-classifier using the outputs of the individual classifiers [26, 276]. It is a scheme for minimizing the prediction errors of one or more learners. Efficient integration by the meta-classifier at the second layer is important as well as the performance of the individual models at the first layer. In this subsection, I describe several NLP studies using SG for better performance.

Stacked generalization is different from weighted majority voting [155] or cascading learning [93]. Weighted majority voting generally determines a voting weight for each individual classifier, while stacked generalization can assign different weights to different types of predictions. Training in cascading learning requires multiple rounds of learning,

while stacked generalization typically consists of just two stages. Also, cascading learning does not need multiple base learners.

Ting and Witten [252] showed that SG using confidence scores from the predictions of multiple classifiers obtained better results than the individual systems. Džeroski and Zeno [75] showed good performance for SG on a collection of 21 datasets from the UCI Repository of machine learning databases [186]. Nivre and McDonald [194] applied SG to dependency parsing by integrating two different models (graph-based models and transition-based models). Rajani et al. [205] used SG for an English Slot Filling task [243, 244]. In addition to the outputs and confidence scores of each model, they developed new features to quantify the reliability of individual models.

Recently, some research has used stacked generalization in the bioinformatics domain. Wang et al. [273] used SG with two base learners for predicting membrane protein types. Netzer et al. [184] applied SG to identify "breath gas markers" and reported improved classification accuracy. Li et al. [148] combined multiple sequence tagging modes by union, intersection, voting and SG methods. Their experimental results showed that the SG approach was more effective than other ensemble methods. For extracting drug-drug interactions, He at al. [101] presented an SG-based approach combining three individual kernels: feature-based, graph, and tree kernels. In clinical NLP, Kilicoglu et al. [127] used stacked generalization for document-level classification to identify rigorous, clinically relevant studies.

Although SG is another attractive ensemble technique, it has been applied less than voting ensembles because it needs to train a meta-classifier that uses the outputs over the training data. I propose an SG ensemble to combine multiple components for better performance, especially more precise extraction of medical concepts in specialty areas. To our knowledge, this is the first work that combines broad medical components and specialty area components in an SG ensemble for MCE.

2.4.3 Domain Adaptation

Our work is also related to supervised domain adaptation, which can be applied when some labeled data for the target domain is available. Many algorithms for efficient domain adaptation have been proposed, and domain adaptation-based models have been shown

to improve performance for some tasks when limited annotated data is available for the target domain. For named entity detection, Florian et al. [86] introduced a method that builds on a source domain model and uses its predictions as features to train the target domain model. Chelba and Acero [37] used the feature weights of the source domain model as a Gaussian prior for initializing each feature in the target domain model. They applied their approach to recover the correct capitalization of uniformly cased text. Foster and Kuhn [87] linearly interpolated source and target domain models for machine translation. Daumé III [63] presented a feature augmentation method that can learn trade-offs between source/target and general feature weights.

Jiang and Zhai [114] proposed several adaptation heuristics with a unified objective function: (1) removing misleading training instances in the source domain, (2) assigning more weights to labeled target instances than labeled source instances, (3) augmenting training instances with target instances with predicted labels. They evaluated the proposed method on three adaptation problems in NLP, POS tagging, NE type classification, and spam filtering. The results showed that their method capturing domain differences explicitly outperforms regular weakly-supervised and supervised learning methods. They also showed that adding more target domain data with high weights is much more useful for improving performance than excluding misleading training examples from the source domain.

Blitzer et al. [20] introduced structural correspondence learning (SCL) to automatically induce correspondences among features from different domains. SCL identifies correspondences among features from different domains by modeling their correlations with pivot features. Their method showed performance gains on part of speech tagging for varying amounts of the source and target training data. Blitzer et al. [19] extended the SCL algorithm to sentiment analysis and also showed how A-distance, a measure of domain similarity, correlates well with the potential for adaptation of a classifier from one domain to another. Xiao and Guo [278] applied the combination of SCL [20] and feature augmentation based [63] domain adaptations to sequence labeling. Their results showed that the proposed domain adaptation method outperforms a number of comparison methods for cross-domain sequence labeling tasks including syntactic chunking and NER.

For comparison to my ensemble methods, I employ the feature augmentation method [63] to combine the data from the source domain (broad medical texts) and the target domain

(specialty notes). The outputs of multiple models trained on both source and target domains are also exploited in voting and SG ensembles.

According to the research discussed above, I can say with fair certainty that ensemble based and domain adaptation approaches have been contributory to various NLP tasks. As pointed out by prior studies [198, 258], the diversity and accuracy of a new component should be considered to make the component more complementary. With these issues in mind, I now shift the emphasis away from combining models to selecting instances to train a new model.

2.5 Exploiting Unlabeled Data

In this section, I discuss the methods that employ labeled and unlabeled data in the learning process. Annotating data is as an expensive job, especially in the biomedical and clinical domains because of the need for domain experts. Consequently, most systems are trained with relatively small amounts of labeled text, even though much larger amounts of unlabeled text are readily available. To build a better model than only with labeled data, considerable studies have been focused on selecting the beneficial instances from unlabeled data. Weakly supervised learning, also called semi-supervised learning, active learning, or combining the two approaches have been applied to improve the performance of several NLP tasks.

To improve medical relation classification, I exploit large amounts of unlabeled clinical texts for self-training, which is a form of weakly supervised learning. However, my new methods select the instances not merely by a classifier's confidence scores, but by using similarity measures to consider representativeness and diversity, which have been critical factors in active learning. In the following subsections, I briefly describe several methods of weakly supervised learning and active learning related to this dissertation research.

2.5.1 Weakly Supervised Learning

Weakly supervised learning has been shown to benefit from training on both labeled and unlabeled data for several NLP tasks including word sense disambiguation [282], web document classification [21], NER [52], noun phrase chunking [199], parsing [163], question answering [206], and relation extraction (discussed in Section 2.3). Self-training is one of the

popular weakly supervised learning methods. Starting with a small set of seed labeled data, a learner uses its own predictions on unlabeled data to retrain the model. The predicted instances that satisfy a selection criterion are collected as new training instances for the next iterations. This procedure of automatically labeling new training examples is performed repeatedly until a stopping criterion is met (e.g., for a fixed number of iterations or until no new instances can be labeled). As seen in Section 2.3, a significant number of studies have demonstrated that relation extraction applying self-training can yield satisfactory results. Depending on the target task, a well-performing classifier is often needed for self-training when the confidence scores produced by the classifier are taken into account for instance selection. Rosenberg et al. [217] also pointed out that the choice of the initial seeds has a large effect on performance. I apply self-training with confidence-based instance selection to compare to my two clustering-based selection methods.

Co-training [21] also starts with a set of labeled data but expands the training data with two (or possibly multiple) classifiers. The classifiers in co-training are iteratively retrained with unlabeled examples collected by the other classifiers. Blum and Mitchell [21] showed that for successful co-training with two separate classifiers, each classifier should already make good classifications, and they should be conditionally independent given the class label. Later, some research [6, 14] showed that co-training with classifiers that are not conditionally independent still can improve performance over a supervised learner. More examples of co-training applications can be found in [47, 52, 120, 169, 187, 199, 222, 270].

NLP researchers also have used graph-based learning methods in which labeled and unlabeled instances are represented as vertices in a connected graph. Label propagation algorithms are one of the major graph-based methods. In the label propagation framework, the information of labeled instances is iteratively propagated to nearby vertices until the labels of unlabeled instances are all determined. The assumption of this approach is that a labeled instance can share the label with any unlabeled instances by propagating the label information to neighboring instances across related edges. Some studies focused on label propagation were reported in [38, 193, 256, 303].

Other weakly supervised learning approaches include transductive learning [117, 118], generative mixture models [67, 68, 94, 192], and distant supervision [54, 174, 281]. For a more comprehensive review of weakly supervised learning, the reader can refer to Zhu's survey

papers [301, 304].

2.5.2 Active Learning

Many NLP researchers may agree that creating a high-quality manually annotated corpus is important but often more complicated than deciding on the learning algorithm. To reduce the laborious manual annotation, active learning has focused on how to collect informative samples from unlabeled data and provide them to the oracle (e.g., a human expert) for labeling. The main interest of active learning is to prioritize the instances to be annotated for classification to select instances that will be most valuable for the classifier to learn from.

Uncertainty Sampling [146] is a popular selection strategy. The most uncertain example is selected in this method. The uncertainty can be measured by prediction confidence, margin score, or entropy [108]. Uncertainty sampling has been applied to several NLP tasks including text classification [145, 146, 255] and sequence labeling [55, 230]. In clinical NLP, Chen et al. [39] applied uncertainty sampling to phenotyping tasks and showed that uncertainty sampling outperformed random sampling.

Query by Committee (QBC) [231] is another active learning strategy which is based on disagreement between multiple learners. The example which the learners most disagree on is selected in this method. Disagreement between learners in the committee (ensemble) can be quantified with entropy or several similarity functions including Kullback-Leibler divergence, or Jensen-Shannon divergence. Examples of QBC applications can be found in [49, 61, 162, 219, 230].

Other instance selection measures such as *Expected Gradient Length* and *Variance Reduction* have been suggested although they may be computationally expensive. Expected Gradient Length [230] selects the instance that causes the maximal change to the current model. In Variance Reduction [219, 289], the instance that minimizes standard error is selected.

The approaches mentioned above rely on the instance selection metric (e.g., uncertainty) and treat each instance from unlabeled data independently. Alternatively, some active learning approaches consider the correlation between instances [91, 189, 240, 300]. These methods often incorporate clustering algorithms to group instances and select the prototype instances in each cluster. For this dissertation research, I similarly apply clustering algorithms

to group instances into clusters. I group both labeled and unlabeled instances together into clusters to get a diverse and balanced set of additional training instances.

In active learning, the oracle is assumed to be never wrong and always can answer the question without any cost. *Proactive learning* [73] was proposed to relax these unrealistic assumptions and focuses on jointly selecting the optimal oracle and instances. There has also been some research [106, 162, 274, 305] to combine active learning and weakly supervised learning.

In this subsection, I introduced some studies of active learning and explained how my clustering-based instance selection methods were motivated by active learning. For a more comprehensive review of active learning, the reader can refer to [228, 229, 239, 272].

CHAPTER 3

MEDICAL CONCEPT EXTRACTION ACROSS DIFFERENT TYPES OF CLINICAL NOTES

In this chapter, I focus on extracting medical concepts in EHRs. I investigate methods for creating medical concept extraction (MCE) systems that will perform well on specialty area notes. For this research, I created three new text corpora consisting of medical notes from three specialty areas: cardiology, neurology, and orthopedics. I present an analysis of how they differ in content (semantic concepts and formatting) from each other and i2b2 medical notes. I will refer to the i2b2 notes as **broad** medical texts because they describe a patient's overall care and their content can cover a diverse set of topics cutting across many areas of medicine.

I examine a variety of information extraction (IE) models and evaluate their performance on all of these data sets. I investigate how MCE models perform on specialty notes (1) when trained on a broad medical corpus, (2) when trained on the same type of specialty data, and (3) when trained on a combination of both broad medical and specialty data.

Another goal of this research is to explore the use of stacked generalization learning for the medical concept extraction task. Stacked generalization provides an easily extensible and adaptable framework for benefiting from an ensemble of extraction models. I explore Voting and Stacked Learning ensembles to combine multiple MCE models and conclude that they can be beneficial by exploiting (1) multiple models that use different extraction techniques, and (2) multiple models trained with specialty area data as well as multiple models trained with broad medical data.

3.1 Data Sets and Annotated Concepts

My research starts with the medical concept extraction (MCE) task defined for the 2010 i2b2 Challenge [263]. This task involves extracting three types of medical concepts: *Problems*

(e.g., diseases and symptoms), *Treatments* (e.g., medications and procedures), and *Tests*.

The 2010 i2b2 corpus consists of 349 training documents and 477 test documents manually annotated by medical professionals. This test set contains 45,009 annotated medical concepts. For this research, I created new text collections representing three specialized areas of medicine: cardiology, neurology, and orthopedics. 200 clinical notes from the BLUlab corpus for each specialty area were annotated.[1]Doing-Harris et al. [72] divided the BLUlab corpus into nine specialty groups and their specialty area annotations were used for this research. Each specialty data set consists of different subtypes of notes. Table 3.1 shows the five most prevalent subtypes in each specialty data set.

For this research, two people with medical expertise manually annotated the specialty notes using the 2010 i2b2 Challenge guidelines [3]. One annotator had previously annotated data for the official 2010 i2b2 Challenge data, and the other annotator had equivalent medical knowledge.[2]After joint discussion on 10 practice documents, I measured their interannotator agreement on 40 new documents (one-third for each area) annotated by both annotators during the pilot phase using Cohen's kappa [36], and their IAA was $\kappa = .67$. Table 3.2 shows interannotator agreement for each batch (10 documents per batch).

Each of the annotators then labeled 100 new documents for each specialty area, producing a total of 600 annotated specialty area texts. These texts contain 17,783 annotated concepts for cardiology, 11,019 concepts for neurology, and 12,769 concepts for orthopedics.

Table 3.3 shows the number of annotated concepts of each type in the i2b2 test data and my three specialty data sets, as well as the average number of concepts per document. For example, the Cardiology data contains 7,474 *Problem* concepts, and the average number of *Problem* concepts per text is 37, which is similar to the i2b2 data (39). However, the Neurology and Orthopedics data sets contain only 25 *Problem* concepts per document, on average. For *Treatment* concepts, the neurology notes contain fewer concepts than the i2b2 data but the orthopedics notes contain more. The prevalence of *Test* concepts varies greatly: the i2b2 and cardiology texts have many *Test* concepts per document, but they are much

[1]The BLUlab corpus is a collection of de-identified clinical notes drawn from multiple clinical settings at the University of Pittsburgh. The dataset was available for research to investigators with local Institutional Review Board approval, but unfortunately the University of Pittsburgh has withdrawn the corpus for new studies. However interested researchers can collaborate with previously approved sites.

[2]They are both registered nurses (RNs).

Table 3.1: Five Most Prevalent Note Subtypes in Each Specialty Area Data Set

Data	**Note subtypes**
Cardiology	Cardiology (surgery) discharge summary
	Cardiology (surgery) consultation report
	Cardiology operative report
	Cardiology history and physical examination
	Angio report
Neurology	Neurosurgery discharge summary
	Neurosurgery transfer summary
	Neurology consultation report
	Neurology history and physical examination
	Neurosurgery death summary
Orthopedics	Orthopedic (surgery) operative report
	Trauma discharge summary
	Orthopedic (surgery) discharge summary
	Orthopedic surgery transfer summary
	Orthopedics consultation report

Table 3.2: Cohen's kappa for Each Batch of 10 Documents

Batch	κ
1	.64
2	.69
3	.67
4	.67

Table 3.3: The Numbers of Concepts in Each Data Set

Categories	**i2b2 Test**		**Cardiology**		**Neurology**		**Orthopedics**	
	Total	**Avg**	**Total**	**Avg**	**Total**	**Avg**	**Total**	**Avg**
Problem	18,550	39	7,474	37	4,971	25	5,022	25
Treatment	13,560	28	5,706	29	3,815	19	6,494	33
Test	12,899	27	4,603	23	2,233	11	1,253	6
All Concepts	45,009	94	17,783	89	11,019	55	12,769	64
# Sentences	45,052	94	21,255	106	15,339	77	16,855	84
# Documents	477		200		200		200	

less common in the neurology notes (11 per text) and orthopedics notes (6 per text).

The fifth row of Table 3.3 compares the number of sentences in the data sets. The i2b2 test data contains 45,052 sentences (94 per file, on average). The cardiology notes were generally longer with 106 sentences per text, while the neurology and orthopedics notes were generally shorter.

I also examined, qualitatively, the types of sections in each data set to gain more insight about content differences between specialist notes and the more general i2b2 notes (broad medical texts). Table 3.4 shows the five most frequent section titles in each data set. Many section titles, such as "Hospital course," are common across all of the data sets. However, I found section titles that are much more frequent in some types of specialty area notes. For example, sections related to "Procedures" and "Operations" occurred most frequently in orthopedics notes. "Consultation" sections were common in the cardiology notes but rare in the i2b2 notes. Appendix A provides more section headers that frequently appear in each dataset.

Although some of the same section titles occur in both broad medical notes and specialty notes, their contents can differ. For example, in the sections titled "Procedures", orthopedics notes typically contain more detailed information than discharge summaries. Appendix B illustrates specialty notes that are similar to the ones in our collection.

3.2 Concept Extraction Models

I developed four types of concept extraction models that use a diverse set of extraction techniques. I will first describe each model and then present my ensemble-based learning framework.

3.2.1 MetaMap

I used a widely-used knowledge-based system called MetaMap [10]. MetaMap is a rule-based program that assigns UMLS Metathesaurus semantic concepts to phrases in natural language text. Unlike my other IE systems, MetaMap is not trained with machine learning, so it is not dependent on training data. Instead, MetaMap is a complementary resource that contains a tremendous amount of external medical knowledge.

I encountered one issue with using this resource for this task. MetaMap can assign a large

Table 3.4: Five Most Frequent Section Titles in Each Data Set

Data	**Section titles**
i2b2 Test	HOSPITAL COURSE
	HISTORY OF PRESENT ILLNESS
	PHYSICAL EXAMINATION
	PAST MEDICAL HISTORY
	ALLERGIES
Cardiology	PHYSICAL EXAMINATION
	ALLERGIES
	PAST MEDICAL HISTORY
	SOCIAL HISTORY
	HISTORY OF PRESENT ILLNESS
Neurology	HOSPITAL COURSE
	REASON FOR ADMISSION
	HISTORY OF PRESENT ILLNESS
	DISCHARGE MEDICATIONS
	DISCHARGE INSTRUCTIONS
Orthopedics	HOSPITAL COURSE
	PROCEDURES
	DISCHARGE INSTRUCTIONS
	DESCRIPTION OF OPERATION
	COMPLICATIONS

set of semantic categories, many of which are not relevant to the i2b2 concept extraction task. However, it is not obvious how to optimally align the MetaMap semantic categories with our task's semantic categories because their coverage can substantially differ.

Therefore I built a statistical model based on the concepts that MetaMap detected in the training data. I collected all of MetaMap's findings in the training data, aligned them with the gold standard medical concepts, and calculated the probability of each MetaMap semantic category mapping to each of this task's three concept types (*Problem*, *Treatment*, and *Test*). Then I assigned a MetaMap semantic type to one of our concept types if the semantic type is ranked among the top 30% of semantic types based on Probability(concept_type | sem_type).[3] For example, "sosy" ("Sign or Symptom" in MetaMap) was mapped to the *Problem* concept type because it had a high probability of being aligned with labeled problems in the data set (i.e., Prob(*Problem* | *sosy*) was high). Table 3.5 shows the semantic types that we ultimately used for concept extraction. Please refer to Appendix C for the mapping between abbreviations and the full semantic type names. I used MetaMap 2013v2 with the 2013AB NLM relaxed database.[4]

3.2.2 Rules

I used the training data to automatically create simple rules. The idea is to exploit the training data to create a simple rule-based system without any manual effort.

First, I computed Prob(concept | word) and Prob(concept_type | word) for each word in the training data, where concept_type = {*Treatment*, *Test*, *Problem*}. Note that Prob(concept | word) is the sum of Prob(*Treatment* | word), Prob(*Test* | word), and Prob(*Problem* | word). Next, I selected words that had frequency ≥ 3 and Prob(concept | word) $\geq .80$. For each selected word, I chose the concept type with the highest probability and created a rule (e.g., *diabetes* $\rightarrow$ *Problem*).

Given a new text, I then found all words that matched a rule and labeled them as concepts using the concept type assigned by the rule. When two or more labeled words were contiguous, I treated them as a single concept. For multiword concepts, I calculated the

[3]Higher F_1 score was achieved with the top 30% of semantic types on the i2b2 training data than with other values (top 10%, 20%, 40%, etc.).

[4]I used the following MetaMap options with the following arguments:
`-C -V NLM -y -i -g --composite phrases 3 --sldi`

Table 3.5: MetaMap Semantic Types Used for Medical Concept Extraction

Category	**MetaMap semantic types**
Problem	acab, anab, bact, celf, cgab, chvf, dsyn, inpo, mobd, neop, nnon, orgm, patf, sosy
Treatment	antb, carb, horm, medd, nsba, opco, orch, phsu, sbst, strd, topp, vita
Test	biof, bird, cell, chvs, diap, enzy, euka, lbpr, lbtr, mbrt, moft, phsf, tisu

average P(concept_type | word) across the words in the concept. The concept type with the highest average probability was assigned to the concept.

3.2.3 Contextual Classifier (SVM)

I created a supervised learning classifier with contextual features. I applied the Stanford CoreNLP tool [157] to our data sets for tokenization, lemmatization, part-of-speech (POS) tagging, and named entity recognition (NER). I trained a support vector machine (SVM) classifier with a linear kernel for multiclass classification using the LIBLINEAR (Library for Large Linear Classification) software package [80].

I reformatted the training data with IOB tags (B: at the beginning, I: inside, or O: outside of a concept). I defined features for the targeted word's lexical string, lemma, POS tag, affix(es), orthographic features (e.g., Alphanumeric, HasDigit), named entity tag, and pairwise combinations of these features. The feature set used to create the SVM model, as well as the CRF models described later, is as follows.

- Word Features:
  - w_0 (current word)
  - w_{-1} (previous word), w_1 (following word)
  - w_{-2} (second previous word), w_2 (second following word)
- Bi-grams of Words:
  - [w_{-1}, w_0] (bi-gram of previous word and current word), [w_{-2}, w_{-1}]
  - [w_0, w_1], [w_1, w_2]
- Lemmas Features:
  - l_{-3} (lemma of third previous word), l_{-2}, l_{-1}, l_1, l_2, l_3
- Affixes Features:
  - Prefixes and suffixes of current word, up to a length of 5. For example, the word "*disease*" would have features "*d*", "*di*", "*dis*", "*dise*", and "*disea*" for prefixes, and "*e*", "*se*", "*ase*", "*ease*", and "*sease*" for suffixes.

- Orthographic Features:
  - 15 features based on regular expressions for w_0, w_{-1}, w_1. I used the orthographic features defined in [164] with a slight modification. For example, "`[A-Z].*`" is a regular expression to capture whether the first letter of a word is capitalized (Init Caps).
- POS Features:
  - p_0 (POS tag of current word), p_{-2}, p_{-1}, p_1, p_2
- Bi-grams of POS tags:
  - $[p_{-1}, p_0]$ (POS tags of previous word and current word), $[p_{-2}, p_{-1}]$
  - $[p_0, p_1]$, $[p_1, p_2]$
- Lemma + POS:
  - $[l_0, p_0]$ (lemma of current word and POS tag of current word)
- NER class:
  - n_0 (e.g., PERSON, LOCATION, DATE, TIME)

I set the cost parameter to be $c = 0.1$ (one of LIBLINEAR's parameters) after experimenting with different values by performing 10-fold cross validation on the training set.

3.2.4 Sequential Classifier (CRF)

I trained several sequential taggers using linear chain conditional random fields (CRF) supervised learning models. In contrast to the contextual classifier mentioned above, the CRF classifiers use a structured learning algorithm that explicitly models transition probabilities from one word to the next. The CRF models used the same feature set as the SVM models.

For each data set, I implemented two different variations of sequential classifiers. I trained CRF classifiers with both forward (CRF-fwd) and backward tagging (by reversing the sequences of words) (CRF-rev) [84, 139]. As a result, each medical concept had different IOB representations. For example, the IOB tags of "positive lymph nodes" by forward

and backward tagging were "*positive/B-problem lymph/I-problem nodes/I-problem*" and "*positive/I-problem lymph/I-problem nodes/B-problem*", respectively.

I used Wapiti [142], which is a simple and fast discriminative sequence labeling toolkit, to train the sequential models. As with the SVM, 10-fold cross validation was performed on the training set to tune Wapiti's CRF algorithm parameters. I set the size of the interval for the stopping criterion to be $e = 0.001$. For regularization, $L1$ and $L2$ penalties were set to 0.005 and 0.4, respectively.

3.3 Ensemble Methods

I explored two types of ensemble architectures that use the MCE methods described previously as components of the ensemble. I created a voting ensemble, as a simple but often effective ensemble method, and a stacked generalization ensemble, which trains a meta-classifier with features derived from the outputs of its component models.

Three different types of ensembles were created for both architectures: (1) ensembles consisting of MCE models trained on the broad medical data, (2) MCE models trained on specialty data, and (3) a mix of MCE models, some trained from broad medical data and others trained from specialty data. For comparison to (3), I also trained a CRF model using the union of the broad medical data and specialty data. I investigated how they perform differently and which advantages my ensembles can offer over the union approach.

3.3.1 Voting Ensemble Method

This ensemble collects the phrases labeled by a set of MCE components and outputs all phrases that received at least three votes (i.e., were labeled by at least three components). I found that three votes worked consistently well. However, that is a disadvantage of voting ensemble: determining the voting threshold can be difficult. The empirical discussion will be examined in Section 3.4.

In the case of overlapping phrases, I choose the one with the highest confidence, based on the normalized confidence scores of the MCE models. For each MCE model, each confidence score was divided by the highest score produced by that model for normalization.

3.3.2 Stacked Generalization Method

I created a meta-classifier by training an SVM classifier with a linear kernel based on the predictions from the individual classifiers. Figure 3.1 shows the architecture of the stacked learning ensemble.

First, to create training instances for a document, all of the concept predictions from the individual IE models are collected. I then use a variety of features to consider the degree of agreement and consistency between the IE models. For each concept predicted by an IE model, it is compared with all other distinct concepts predicted in the **same** sentence. For each pair of concepts, the following eight matching criteria are applied to create eight features:

- If the text spans match
- If the text spans partially match (any word overlap)
- If the text spans match and concept types match
- If the text spans partially match and the concept types match
- If the text spans have the same start position
- If the text spans have same end position
- If one text span subsumes the other
- If one text spans is subsumed by the other

For example, with two concepts from a sentence "*avoid overdosing on insulin before exercise*",

C_1: "*overdosing on insulin*" (*Problem*)
C_2: "*overdosing*" (*Treatment*)

"*No*" is assigned to '*if the text spans match*'. "*Yes*" and "*No*" to '*if the text spans partially match*' and '*if the text spans partially match and the concept types match*, respectively.

Features are also defined to count how many different models produced a predicted concept, and features are defined for predictions produced by just a single model (indicating which model produced the predicted concept). To make sure that the meta-classifier was

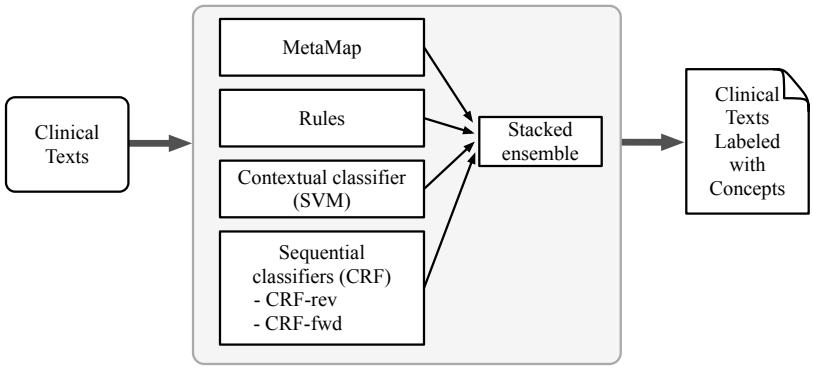


Figure 3.1: Architecture for a Stacked Learning Ensemble

neither trained nor applied to the same document, I performed a procedure similar to 10-fold cross validation on the training set to obtain predictions for each classifier.

3.4 Evaluation of MCE Models and Ensemble Methods

The input for this task is clinical texts. Medical concept annotations from the gold standard data are used to train the clssifiers. Figure 3.2 illustrates a sample text with gold concept annotations.

The text files are formatted to have one sentence per line. The concept annotation file contains medical concepts, one concept per line. The concept annotation format specified in the 2010 i2b2 Challenge Annotation File Formatting [1] is as follows:

`c="concept text" offset||t="concept type"`

where

`c` represents a mention of a concept.

`concept text` is filled in with the actual text from the text file.

`offset` represents the beginning and end line and word numbers that span the concept.

`t` represents the semantic type of the concept mentioned.

`concept type` is replaced with *problem*, *treatment*, or *test*.

An offset is formatted as the line number followed by a colon followed by the word number. The starting offset and ending offset are separated by a space.

3.4.1 Evaluation Metrics

I used three metrics to evaluate medical concept extraction: recall, precision, and F_1 score. From the counts of true positives (TP), false negatives (FN), and false positives (FP), *Recall* and *Precision* can be computed as follows:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{3.1}$$

The F_1 score is defined by taking the harmonic mean of recall and precision [266]:

$$F_1 = 2\,\frac{precision \times recall}{precision + recall} = \frac{2\,TP}{2\,TP + FP + FN} \tag{3.2}$$

Each metric was microaveraged across all concepts in the evaluation set. For all

```
1    The patient is a 64-year-old male with a long standing history of
.    peripheral vascular disease who has had multiple vascular procedures
.    in the past including a fem-fem bypass , a left fem pop as well as
.    bilateral TMAs and a right fem pop bypass who presents with a
.    nonhealing wound of his left TMA stump as well as a pretibial ulcer
.    that is down to the bone .
2    The patient was admitted to obtain adequate pain control and to have
.    an MRI / MRA to evaluate any possible bypass procedures that could be
.    performed .
```

(a) A Sample Text (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc3.txt)

```
c="peripheral vascular disease" 1:12 1:14||t="problem"
c="multiple vascular procedures" 1:18 1:20||t="treatment"
c="a fem-fem bypass" 1:25 1:27||t="treatment"
c="a left fem pop" 1:29 1:32||t="treatment"
c="bilateral tmas" 1:36 1:37||t="treatment"
c="a right fem pop bypass" 1:39 1:43||t="treatment"
c="a nonhealing wound of his left tma stump" 1:47 1:54||t="problem"
c="a pretibial ulcer" 1:58 1:60||t="problem"
c="adequate pain control" 2:6 2:8||t="treatment"
c="an mri / mra" 2:12 2:15||t="test"
c="bypass procedures" 2:20 2:21||t="treatment"
```

(b) A Sample Concept Annotations (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc3.com)

Figure 3.2: A Sample Text With Concept Annotations

experiments shown in this chapter, a labeled phrase was scored as correct if it was assigned the correct concept type and its text span exactly matched the gold standard text span, disregarding articles and possessive pronouns (e.g., "*his*"). The reader can refer to Appendix D for additional partial match results.

3.4.2 Statistical Significance Testing

I used a two-tailed paired t test across the F_1 scores for all test documents. Each F_1 score was calculated for each document and then averaged across all test documents. The significance level, α, was set at 5%.

3.4.3 Performance of Individual MCE Models

I conducted an extensive set of experiments to evaluate the performance of each individual MCE model and voting and stacked generalization ensembles. I also experimented with models trained using the broad medical (**i2b2**) texts, using our specialty area texts, and using a mixture of both.

I evaluated performance using the i2b2 training and test sets as well as our three sets of specialty area notes: cardiology, neurology, and orthopedics (described in Section 3.1). The specialty area models (**Sp**) were trained and evaluated using 10-fold cross validation on my specialty notes data. First, I present experimental results for each individual MCE model. Table 3.6 shows the performance of each MCE model based on Recall, Precision, and F_1 score.

MetaMap: The **MetaMap** row shows low scores for MetaMap on all data sets. As explained before, MetaMap suffers from boundary mismatch errors due to the difference between the i2b2 annotations and MetaMap's concept boundary definitions. I also observed that MetaMap often did not recognize acronyms and abbreviations in the clinical notes.

Rules: The **Rules (i2b2)** rows show results for the simple rules harvested from the i2b2 training data. Not surprisingly, these rules performed better on the i2b2 test set than on the specialty notes, but the scores were low across the board. The **Rules (Sp)** rows show results (averaged during cross-validation) for the rules harvested from the training folds for a specialty area and evaluated on the test folds for the same specialty area. These rules also performed poorly.

SVM: The machine learning classifiers performed substantially better. The **SVM (i2b2)**

Table 3.6: Results of Individual MCE Models

Model	Recall	Precision	F_1 score
i2b2 (Broad Medical) Evaluation			
MetaMap	36.0	47.3	40.9
Rules (i2b2)	38.5	48.4	42.9
SVM (i2b2)	80.6	76.9	78.7
CRF-fwd (i2b2)	81.4	86.1	83.7
CRF-rev (i2b2)	**82.3**	**86.4**	**84.3**
CRF-rev ($i2b2_{180}$)	78.7	84.2	81.4
Cardiology Specialty Area Evaluation			
MetaMap	31.1	40.0	35.0
Rules (i2b2)	33.1	37.9	35.3
Rules (Sp)	32.6	38.6	35.3
SVM (i2b2)	64.5	59.4	61.8
SVM (Sp)	65.5	59.4	62.3
CRF-fwd (i2b2)	65.2	67.9	66.5
CRF-rev (i2b2)	65.8	68.0	66.9
CRF-rev ($i2b2_{180}$)	63.3	66.5	64.9
CRF-fwd (Sp)	63.8	69.3	66.4
CRF-rev (Sp)	65.2	69.1	67.1
CRF-rev (i2b2+Sp)	**68.7**	**70.3**	**69.5**
Neurology Specialty Area Evaluation			
MetaMap	29.4	34.6	31.8
Rules (i2b2)	29.2	35.3	32.0
Rules (Sp)	30.9	33.0	31.9
SVM (i2b2)	59.4	55.7	57.5
SVM (Sp)	60.2	53.8	56.8
CRF-fwd (i2b2)	61.3	65.8	63.5
CRF-rev (i2b2)	61.7	65.7	63.6
CRF-rev ($i2b2_{180}$)	59.6	64.8	62.1
CRF-fwd (Sp)	59.2	64.6	61.8
CRF-rev (Sp)	60.5	64.6	62.5
CRF-rev (i2b2+Sp)	**64.6**	**66.8**	**65.7**
Orthopedics Specialty Area Evaluation			
MetaMap	22.6	26.3	24.3
Rules (i2b2)	21.4	26.2	23.5
Rules (Sp)	26.8	27.9	27.3
SVM (i2b2)	45.7	41.6	43.5
SVM (Sp)	56.6	49.1	52.6
CRF-fwd (i2b2)	47.4	56.3	51.5
CRF-rev (i2b2)	48.2	55.8	51.7
CRF-rev ($i2b2_{180}$)	44.8	53.3	48.7
CRF-fwd (Sp)	55.4	62.3	58.6
CRF-rev (Sp)	56.0	60.6	58.2
CRF-rev (i2b2+Sp)	**59.3**	**62.5**	**60.9**

row shows results for the SVM model trained on i2b2 data, which produced an F_1 score of 78.7% on the i2b2 test set but substantially lower F_1 scores on the specialty datasets. The **SVM (Sp)** row shows results for the SVMs trained and tested on specialty area data. Performance substantially improved on the Orthopedics notes (from 43.5% to 52.6% F_1 score) but did not change much for the other specialty areas.

CRF: Both the CRF-fwd and CRF-rev models trained on i2b2 data performed better than the SVM models. The **CRF-fwd (Sp)** and **CRF-rev (Sp)** rows show results for CRF models trained on specialty area data. Performance on the Cardiology and Neurology notes was similar when trained on specialty (Sp) data, but performance on the Orthopedics notes substantially improved.

Since the i2b2 training data is much larger than my specialty area training data, I performed another experiment using only 180 randomly selected i2b2 training texts, to match the amount of specialty area training data (under 10-fold cross-validation, each fold trains with 180 documents). The CRF-rev models performed a little better than the CRF-fwd models, so I conducted this experiment only with the CRF-rev model (shown as **CRF-rev (i2b2$_{180}$)**). The performance of these models is lower than when using all of the i2b2 training data as one would expect. More importantly, these experiments demonstrate that training on specialty area data consistently performs better than training on i2b2 data when using comparable amounts of training data.

The **CRF-rev (i2b2+Sp)** row shows the results for training the CRF-rev model using the union of the i2b2 and specialty area data. Performance improved for all three specialty areas by training with the combined data sets (i.e. merging two datasets). The broad i2b2 data clearly provides added value. This CRF-rev model (CRF-rev (i2b2+Sp)) obtained the best results among the individual MCE models. However, the F_1 scores for the three specialty areas range from 60.9% to 69.5%, which is substantially lower than the 84.3% F_1 score achieved for the i2b2 test set.

3.4.4 Performance of Voting and Stacked Ensembles

I also evaluated the performance of the voting and stacked ensemble architectures, which were populated with five types of MCE components: Rules, MetaMap, SVM, CRF-fwd, and CRF-rev models. For both the voting and stacked architectures, I created three different

types of ensembles: (1) i2b2 ensembles consisting of MCE models trained on the i2b2 data, (2) Sp ensembles consisting of MCE models trained on specialty data, and (3) i2b2+Sp ensembles consisting of MCE models trained with i2b2 data and MCE models trained with specialty data. Consequently, the i2b2+Sp ensembles include nine different classifiers (two models each of Rules, SVM, CRF-fwd, CRF-rev, and one MetaMap model, because it does not use training data).

Table 3.7 shows the performance of these ensembles, as well as the *EasyAdapt* domain adaptation method [63], which I implemented as another point of comparison. For EasyAdapt, I used a CRF-rev classifier with the feature set augmented for broad medical (i2b2) notes as the source domain and specialty area notes as the target domain. For the sake of comparison, the first row of Table 3.7 displays again the results obtained for the best individual MCE model from Table 3.6, which was the CRF-rev classifier trained with both i2b2 and specialty data. Comparing the first two rows, the readers see that training a CRF-rev model with combined i2b2 and specialty area data outperforms the domain adaptation model on all three data sets.

Among the voting ensembles, the i2b2+Sp ensemble produced the best F_1 scores. The voting ensemble trained only on specialty notes (Sp) produced much higher precision than the CRF-rev model. A voting ensemble appears to be an effective way to improve precision on specialty notes when a limited amount of annotated specialty data is available, although with some cost to recall.

The voting threshold is a key parameter for voting ensembles that can dramatically affect performance. The voting threshold can serve as a recall/precision knob to obtain different trade-offs between recall and precision. In Figure 3.3, I show voting (i2b2+Sp) results for cardiology with voting thresholds ranging from two to eight. The curves show that precision increases as the threshold gets higher, but recall drops simultaneously. When the voting threshold exceeds six, recall drops precipitously.

For stacked learning, every stacked ensemble outperformed its corresponding voting ensemble. The best stacked ensemble (i2b2+Sp) included MCE models trained on i2b2 data as well as MCE models trained on specialty data, producing slightly higher F_1 scores than the CRF-rev models for all three specialty areas. Using a paired t test to measure statistical significance, the F_1 score performance of the i2b2+Sp Stacked ensemble is significantly

Table 3.7: Results of Ensemble Methods

Model	Cardiology			Neurology			Orthopedics		
	Rec	**Pre**	**F**	**Rec**	**Pre**	**F**	**Rec**	**Pre**	**F**
CRF-rev (i2b2+Sp)	68.7	70.3	69.5	**64.6**	66.8	65.7	**59.3**	62.5	60.9
EasyAdapt	66.1	69.5	67.8	62.0	65.4	63.7	57.7	62.0	59.8
Voting (i2b2)	61.0	73.0	66.5	56.2	70.4	62.5	40.7	64.2	49.8
Voting (Sp)	58.3	**77.8**	66.7	52.9	**74.3**	61.8	47.3	**73.0**	57.4
Voting (i2b2+Sp)	**69.4**	69.8	69.6	64.5	66.0	65.3	56.6	62.5	59.4
Stacked (i2b2)	65.7	69.0	67.3	61.8	66.9	64.3	47.8	57.6	52.3
Stacked (Sp)	63.4	73.9	68.2	57.4	70.9	63.4	52.3	70.2	60.0
Stacked (i2b2+Sp)	66.0	75.1	**70.2**	61.5	72.4	**66.5**	54.6	70.8	**61.6**

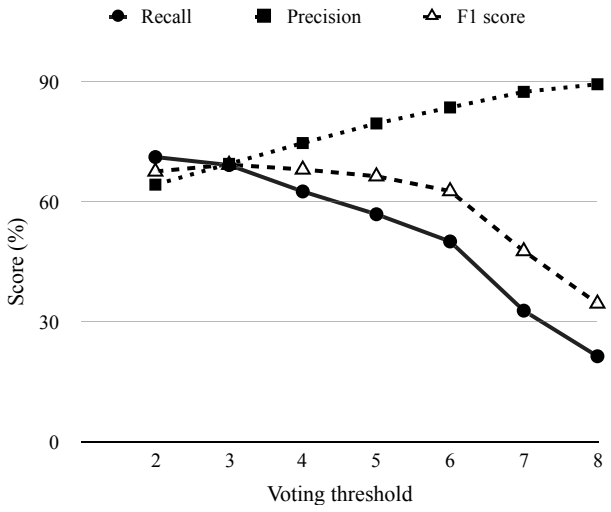


Figure 3.3: Results of the Voting Ensemble for Varying Voting Thresholds (Cardiology)

better than EasyAdapt and all of the voting ensembles at the $p < .05$ significance level, but not significantly better than the CRF-rev (i2b2+Sp) model.

However, the results show that the stacked ensemble produces higher precision than the CRF-rev model (70% → 75% for cardiology; 67% → 72% for neurology; 63% → 71% for orthopedics), with correspondingly smaller decreases in recall (69% → 66% for cardiology; 65% → 62% for neurology; 59% → 55% for orthopedics). My conclusion is that the stacked ensembles consistently produce MCE models with a favorable recall/precision trade-off.

I performed ablation tests for both the voting and stacked generalization ensembles to evaluate the impact of each IE model on the ensembles. An ablated ensemble was tested by removing a single model from the ensemble. Table 3.8 shows the F_1 score for each ablated ensemble and the difference from the F_1 score of the full Stacked (i2b2+Sp) ensemble. I only report the result of ensembles for Cardiology.

Every IE model except MetaMap, Rules (i2b2), and Rules (Sp) contributed to the performance of the voting ensemble. Adding any CRF model or SVM (Sp) resulted in better performance of the stacked ensemble. For the voting ensemble, the F_1 score dropped the most when the CRF-rev (Sp) model was removed. For stacked generalization, removing the SVM (Sp) model had the biggest impact.

3.4.5 Practical Issues

I measured the time for each MCE model to extract medical concepts from the test set. Table 3.9 shows the times that four different MCE models and the stacked ensemble spent with the Cardiology data (200 text files). Two stacked ensembles were measured: the full stacked (i2b2+Sp) ensemble that includes nine different classifiers (as explained in the previous subsection) and an ablated ensemble without MetaMap. The time was averaged per document. All measurements were performed on a MacBook Pro with a 2.5 GHz Intel Core i7 processor and 16 GB of memory.

All individual models except MetaMap and the stacked ensemble recognized the medical concepts in less than 100 milliseconds. MetaMap was the bottleneck and spent 1 minute 56 seconds to process each document. Consequently, the stacked (i2b2+Sp) took 1 minute 56.4 seconds.

Table 3.8: The Ablation Tests of Voting and Stacked Ensembles (Cardiology)

Ablated Model	Voting		Stacked	
	$\mathbf{F_1}$ **score**	**Impact**	$\mathbf{F_1}$ **score**	**Impact**
MetaMap	69.6	0.0	70.2	0.0
Rules (i2b2)	70.0	0.4	70.4	0.2
Rules (Sp)	69.9	0.3	70.2	0.0
SVM (i2b2)	69.1	-0.5	70.2	0.0
SVM (Sp)	68.4	-1.2	69.3	-0.9
CRF-fwd (i2b2)	68.7	-0.9	70.1	-0.1
CRF-fwd (Sp)	68.1	-1.5	69.7	-0.5
CRF-rev (i2b2)	68.7	-0.9	70.0	-0.2
CRF-rev (Sp)	68.0	-1.6	69.8	-0.4

Table 3.9: Prediction Time per Document

Model	Time
MetaMap	1m 56s
Rules	0.006s
SVM	0.051s
CRF	0.071s
Stacked (i2b2+Sp)	1m 56.4s
Stacked (i2b2+Sp) without MetaMap	0.4s

m = Minute and s = Second

3.4.6 Discussion and Analysis

The main conclusion of my research is that models trained with a combination of broad medical data and specialty data consistently perform better than models trained on either type of data alone, when the amount of specialty data is limited. I also find that a stacked ensemble consisting of a diverse set of MCE models using different types of extractors achieves overall performance comparable to the best individual classifier in my experiments, but offers two advantages. First, the stacked ensemble yields a recall/precision balance that favors precision, which may benefit applications that place a premium on high precision. Second, the stacked ensemble can be easily augmented with additional components as new resources become available because the meta-classifier automatically learns how to use them simply by retraining the meta-classifier component. In contrast, adding new components to voting ensembles can require a change in voting strategies, and voting ensembles do not provide a way to learn weights to optimally control the influence of different component models.

To demonstrate this advantage over voting, I added a second copy of the Rule (Sp) component as an additional system in the voting (i2b2+Sp) ensemble for the cardiology specialty. Voting between the ten (= 9 + 1) systems using the original threshold of three dropped the F_1 score by -2.6%. Adding a third copy of the Rule (Sp) component (producing 11 component systems) decreased the F_1 score by -6.9% (absolute).

In the same scenarios, the stacked learning ensemble proved to be much more robust, showing almost no change in performance (0.1% with 10 system and 0% with 11 system). Figure 3.4 shows the F_1 scores of voting and stacked ensembles when copies of the Rule (Sp) component are added one by one. The gray-colored curve with triangle dots represents the voting ensemble. It shows the sharpest decline in F_1 scores as each copy of the Rule (Sp) is added. The black-colored curve with circle dots represents the stacked ensemble. The F_1 score increased from 70.2% to 70.4% even with five extra copies of the Rule (Sp) model in the ensemble, demonstrating the robustness of a stacked learning architecture.

Another finding of this research is that performance on all three types of specialty areas is much lower than performance on the broad medical (i2b2) texts. Clearly, there is ample room for improvement for medical concept extraction from specialty area clinical notes and more work is needed on this topic. To better understand the strengths and weakness of

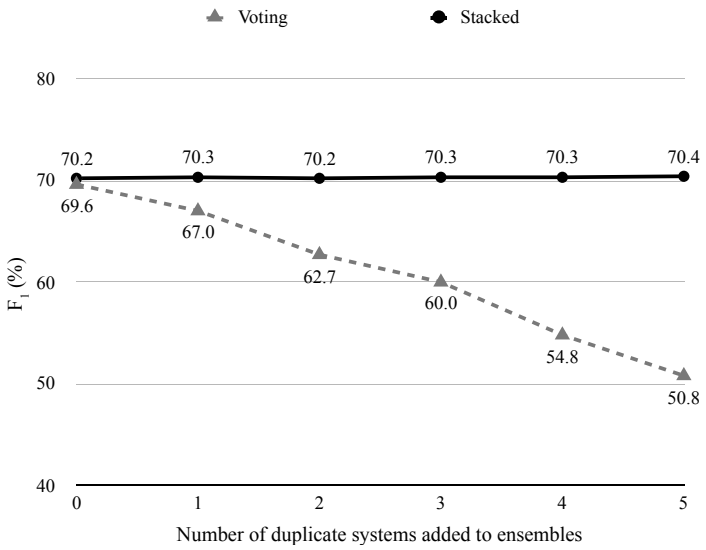


Figure 3.4: Results of the Ensembles by Adding Copies of the Rule (Sp) Component

my models, I manually inspected their output. I observed that my ensemble methods are particularly successful at identifying more accurate concept boundaries than the individual MCE models (e.g., identifying "*severe chest pain*" as a *Problem* concept instead of just "*severe*" or "*chest pain*").

I also analyzed the false negative errors by the CRF-rev models trained with i2b2 data and those trained with specialty data. Table 3.10 shows the results of this manual analysis, which were based on one test fold (20 notes) for each specialty area. The first row of Table 3.10 corresponds to the percentage of false negative errors due to unseen vocabulary. These concepts were misclassified when none of the words in a concept occurred in the training data. For example, the cardiology concepts *thoracoscopy* and *cardioplegia* never appeared in the i2b2 training data. Unseen concepts accounted for roughly the same percentage of errors when training with i2b2 data or specialty data, but note that the i2b2 training set is roughly twice as large as each specialty area training set.

The second row of Table 3.10 corresponds to false negatives for concepts containing at least one seen word and one unseen word. The table shows more false negatives in this category for the models trained with i2b2 data than the models trained with specialty data. For example, for the Treatment concept *aortic crossclamping*, *crossclamping* never appeared in the i2b2 training data but it did appear in the cardiology training data. This type of error was most common in the orthopedics data (51% of the errors), which suggests that the orthopedics notes contain many vocabulary terms that are not present in the i2b2 data.

The third row of Table 3.10 corresponds to false negatives for concepts containing all seen words, but at least one rarely seen word (frequency ≤ 3). For example, in the cardiology data, the concepts *psa data* and *r-wave* were not identified by the i2b2 trained model. The model trained with cardiology data could not extract *nystatin* and *oximeter*, even though they occurred (infrequently) in the cardiology training data.

The last row of Table 3.10 corresponds to false negatives for concepts consisting entirely of words that occurred > 3 times in the training data. Many false negative errors fell into this category. Generally, there were more false negative errors of this type for the models trained with specialty data than those trained with i2b2 data, presumably because the vocabulary is more homogenous in the specialty areas, so more words simply fall into the seen category.

Table 3.10: False Negatives (Percentage) by CRF-rev(i2b2) and CRF-rev(Sp) Models

Error types	**Cardiology**		**Neurology**		**Orthopedics**	
	i2b2	Sp	i2b2	Sp	i2b2	Sp
All unseen	5	6	6	6	8	4
At least one unseen word	31	21	37	21	51	19
At least one word rarely seen	16	17	14	17	14	19
All seen	48	56	43	56	27	58

I observed that many errors were due to incorrect phrase boundaries of medical concepts. For example, only the word *hepatitis* was labeled in the phrase "*hepatitis c*." I also witnessed some tricky errors due to contextual differences in the words surrounding medical concepts. For example, a Treatment concept *lidocaine* is often prescribed for usage on skin ("*treated with lidocaine jelly for pain control*"). However, in the cardiology data, it is usually applied by infiltration ("*Lidocaine 20 cc was infiltrated into the tissues*").

3.5 Improvements to the Broad Medical (i2b2) Concept Extraction

In Chapter 5, I present research on relation classification for the broad medical domain. I used this concept extraction work to identify concepts on unannotated texts for the relation classification work. In this section, I describe further improvements to the broad medical concept extractor to benefit that work for the i2b2 domain.

For weakly supervised learning of medical relation classification (Chapter 5), I used the stacked learning ensemble (Stacked (i2b2)) to identify the medical concepts in the unlabeled data (i.e., MIMIC II Clinical Database [220]). For this preparation, I made some changes in the stacked ensemble.

The first revision I made was the addition of skip-gram features and word embedding features to the feature set of the SVM and CRF classifiers. Word embeddings have contributed to several NLP tasks by providing an alternative representation of information in vector spaces. I used the *Word2Vec* software [171] to perform K-means clustering on the word embeddings. I created 1,000 clusters of semantically related words within the unlabeled data with default parameters of Word2Vec. I used the cluster identifier of each word in a sentence as a feature. Next, in addition to the CRF-fwd and CRF-rev models, I created two versions of CRF classifiers both with MetaMap output as features. The revised i2b2 ensembles include seven different classifiers (MetaMap, Rules, SVM, CRF-fwd, CRF-fwd w/ MetaMap, CRF-rev, and CRF-rev w/ MetaMap).

The concepts annotated by the i2b2 annotation guidelines [3] include modifying articles, pronouns, and prepositional phrases. When applying MetaMap to the training set, I observed that there is a huge difference between the i2b2 annotations and MetaMap's concept boundary definition, especially with respect to articles and pronouns. MetaMap typically excludes

modifying articles, pronouns, and prepositional phrases. For example, for "*a cyst in her kidney*," only "*cyst*" was extracted by MetaMap. Therefore I added a postprocessing step that uses three simple heuristics to adjust concept boundaries to reduce mismatch errors. Although these rules were originally compiled for use with MetaMap, I ultimately decided to apply them to all of the IE models. The three heuristic rules are:

- I include the preceding word contiguous to a detected phrase when the word is a quantifier (e.g., "*some*"), pronoun (e.g., "*her*"), article (e.g., "*the*'), or quantitative value (e.g., "*70%*"). I manually compiled the lists of common quantifiers, pronouns, and articles.

- I include a following word contiguous to a detected phrase when the word is a closed parenthesis (")") and the detected phrase contains an open parenthesis ("(").

- I exclude the last word of a detected phrase when the word is a punctuation mark (e.g., period, comma).

Finally, I added more features to the feature set of the stacked ensemble. I created a feature for the confidence score of each predicted concept: the number of word tokens in a prediction, and whether the prediction contains a conjunction or prepositional phrase. I also created a feature that counts how many times the same phrase was predicted to be a concept in other sentences in the same document.

Table 3.11 shows the performance of other state-of-the-art systems for medical concept extraction alongside the results from the stacked learning ensemble (stacked (i2b2) revised). For comparison with these systems, I report the results of class exact match on the i2b2 test set. In class exact match, both the text span and semantic category must exactly match the reference annotation. I used the i2b2 Challenge evaluation script to compute recall, precision, and F_1 scores. The stacked ensemble produces higher recall and precision than all of the other systems. The F_1 score of the stacked ensemble is comparable to the F_1 score of the best previous system by Tang et al. [247].

3.6 Conclusion

I analyzed the differences in content between broad medical and specialty area notes. Interestingly, orthopedics specialty notes exhibit the most unique language when compared

Table 3.11: Comparison of Other State-of-the-Art Systems With My Stacked Ensemble on the i2b2 Test Set

System	**Recall**	**Precision**	**F_1 score**
de Bruijn et al. [64]	83.6	86.9	85.2
Kang et al. [122]	81.2	83.3	82.2
Tang et al. [247]	84.3	87.4	85.8
Stacked Ensemble	84.4	89.1	86.7

to other specialty notes or broad medical texts.

When a limited amount of annotated specialty area data is available, my research shows that training concept extractors with both broad medical data and specialty area data produces MCE models that achieve better performance on specialty notes than training with either type of data alone. In addition, my research found that a stacked ensemble with mixed domain models, including different types of MCE models as well as models trained on different types of data, achieves good performance and offers some advantages over other approaches. Stacked learning offers the advantage of being able to easily incorporate any set of individual concept extraction components because it automatically learns how to combine their predictions to achieve the best performance.

I found that even though the best individual model (CRF-rev (i2b2+Sp)) and the stacked ensemble produce similar F_1 scores, they exhibit different behaviors with respect to the underlying recall and precision of their output. The stacked ensemble (i2b2+Sp) extracts medical concepts more precisely than the CRF-rev (i2b2+Sp) model trained on the union of the two datasets. Consequently, my results suggest that an individual MCE model may be preferable for applications where recall is more important than precision, while the stacked ensemble may be preferable for applications where precision is more important than recall.

I also observed that MCE performance on specialty texts is substantially lower than state-of-the-art performance on broad medical texts. A promising direction for future work is to explore semisupervised methods to exploit larger collections of unannotated specialty area notes for training.

CHAPTER 4

MEDICAL ASSERTION CLASSIFICATION

In this chapter, I focus on the medical assertions classification task. I present an NLP system that classifies the assertion type of medical problems in clinical notes. Given a medical problem mentioned in a clinical text, an assertion classifier must look at the context and choose the status of how the medical problem pertains to the patient by assigning one of six labels: *present*, *absent*, *hypothetical*, *possible*, *conditional*, or *not associated with the patient*. This task was introduced with medical concept extraction in 2010 for the i2b2 Challenge Shared Tasks [263]. In the i2b2 Challenge data, two types of assertions (*present* and *absent*) are frequently mentioned while the other four types (*hypothetical*, *possible*, *conditional*, and *not associated with the patient*) are less common. Even though the minority classes are not common, they are extremely important to identify accurately (e.g., a medical problem not associated with the patient should not be assigned to the patient).

The assertion information potentially plays a valuable role in medical relation classification which is one of the main tasks in this dissertation research. Each medical problem's assertion has to be annotated when the concept is extracted from unlabeled data and its assertion information is needed to classify the relations associated with the concept. Therefore, this assertion task can act as a bridge between concept extraction and relation classification for successful application of weakly supervised learning to medical relation classification (discussed in Chapter 5).

I approach the assertion classification task as a supervised learning problem. The classifier is given a medical term within a sentence as input and must assign one of the six assertion categories to the medical term based on its surrounding context. First, I describe the each category of assertion with examples provided in the 2010 i2b2 Challenge assertion annotation guidelines [2]. For each category, medical problem concepts in the assertion category appear underlined.

1. PRESENT: problems associated with the patient are present.
    - *patient had a stroke*
    - *the patient experienced the drop in hematocrit*
2. ABSENT: an assertion that the problem does not exist in the patient. It also includes mentions where it is stated that the patient **had** a problem, but no longer does.
    - *patient denies pain*
    - *his dyspnea resolved*
3. POSSIBLE: an assertion that the patient may have a problem, but there is uncertainty.
    - *This is very likely to be an asthma exacerbation*
    - *Doctors suspect an infection of the lungs*
    - *We are unable to determine whether she has leukemia*
4. CONDITIONAL: an assertion that the patient experiences the problem only under certain conditions.
    - *Patient has had increasing dyspnea on exertion*
    - *Penicillin causes a rash*
    - *Patient reports shortness of breath upon climbing stairs*
5. HYPOTHETICAL: an assertion that the patient may develop the medical problems.
    - *If you experience wheezing or shortness of breath*
    - *Ativan 0.25 to 0.5 mg IV q 4 to 6 hours prn anxiety*
6. NOT ASSOCIATED WITH THE PATIENT: an assertion that the medical problem is associated with someone who is not the patient.
    - *Family history of prostate cancer*
    - *Brother had asthma*

I create an SVM classifier with a variety of linguistic features, including lexical, syntactic, lexico-syntactic, contextual, and word embedding features. This study is the subsequent research to Kim et al. [133]. I retrain the SVM model after replacing the preprocessing module and adding new features including word embedding features. In the following sections, I will describe the methods I used for assertion classification and present the experimental results and feature contribution analysis.

4.1 Assertion Classification System

I built a UIMA [82, 83] based assertion classification system with multiple preprocessing components, as depicted in Figure 4.1. The architecture includes a concept importer to parse the concept annotation file with the i2b2 format shown in Figure 3.2, an assertion importer, a section detector, a tokenizer, a lemmatizer, a part-of-speech (POS) tagger, and a context analyzer (local implementation of the ConText algorithm [34]).

I applied the Stanford CoreNLP tool [157] for tokenization, lemmatization, and POS tagging. I used the sentence boundaries that have already been provided by the i2b2 committee (one sentence per line). When a sentence ends with a colon and all words in the sentence are capitalized, the sentence is annotated as a section header.

The assertion classifier uses features extracted by the subcomponents to represent training and test instances. I used the LIBLINEAR software package [80] for multiclass SVM classification with a linear kernel. I tuned the following SVM parameters by 10-fold cross validation on the training set: the cost parameter c was set 0.1 after trials with specified subsets of LIBLINEAR parameters, and the weight of each assertion class was set 0.9, 1.0, 1.0, 1.3, 1.2, and 1.5 for *present*, *absent*, *hypothetical*, *possible*, *conditional*, and *not associated with the patient* respectively. This tuning procedure to minimize the misclassification errors was performed with all features that will be explained in the next section.

4.2 Feature Set Description

I transformed the corpus into a collection of instances to train models for assertion classification. The assertion classifiers used five types of features listed below, which I developed by manually examining the training data:

- Contextual Features:

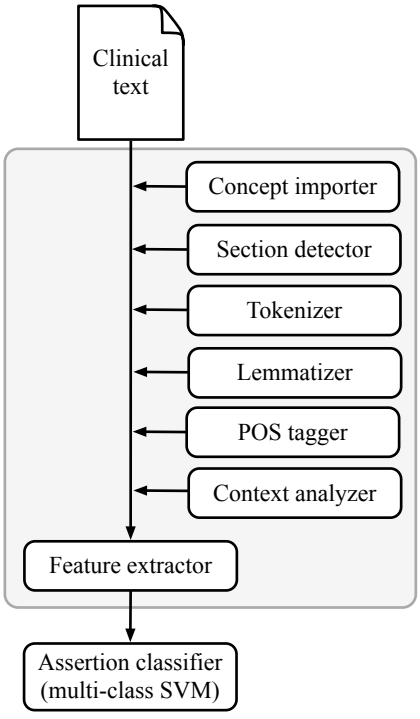


Figure 4.1: System Architecture for Assertion Classification.

- I incorporated the ConText algorithm [34] at the sentence level to detect four contextual properties in the sentence for a medical problem term, C: *absent*(negation), *hypothetical*, *historical*, and *not associated with the patient*. The algorithm assigns one of three values to each contextual property: *true*, *false*, or *possible*. Specifically, four features with three possible values each were defined.

- The ConText algorithm can fail to capture contextual properties because it does not consider the distance between contextual cue (e.g., negation) words and the medical problem term. To reduce incorrect contextual property assignment, I also created a second set of ConText algorithm properties restricted to the six-word context window around the medical problem term (three words on the left and three words on the right).

- The ConText properties were also captured for the medical problem preceding C, if any.

- I identified the section headers that are followed by one or more sentences exclusively containing certain minority assertion classes. Three binary features were defined for *hypothetical*, *conditional*, and *not associated with the patient* to try to improve performance on those specific classes. For instance, according to the assertion annotation guidelines [2], problems associated with allergies were defined as *conditional*. So I created one binary feature that is true if C is underneath the section header containing terms related to allergies (e.g., "*Medication allergies*"). Appendix E provides the complete lists of the section headers identified for *hypothetical*, *conditional*, and *not associated with the patient* classes.

- Lexical Features:

  - For each word contained in the medical term, C, I defined lexical features that include the lowercase version of the word, a canonical form (lemma), and a lemma bi-gram.

  - Lemma of head word in C

  - Lemmas of five words preceding C, and those of five words following it

  - Lowercase version of words in a context window of size five
  - Lemma bi-grams of three words preceding C, and those of three words following it
  - Lemma and lowercase forms of the nearest words before and after C
  - Combination of head lemma and the nearest preceding word lemma, and same combination for the nearest following word lemma
  - Finally, I created a binary feature that is true if the medical term contains a word with a negative prefix.[1]

- Lexico-syntactic Features:
  - I defined features representing words corresponding to several POS in the same sentence as the medical term. The value for each feature is the lemmatized word string. To mitigate the limited window size of lexical features, I defined one feature each for the nearest preceding and following adjective, adverb, preposition, and verb; and one additional preceding adjective, preposition, and verb; and one additional following preposition and verb.
  - I also defined two binary features that check for the presence of a comma or question mark adjacent to the medical term.

- Syntactic Features:
  - POS tags of the words in the medical term
  - POS tags of the five words preceding the medical term and the five words following it

- Word Embedding Features:
  - I used word embedding features derived from the Word2Vec [171] clusters that were computed for medical concept extraction (Section 3.5). I used the cluster identifier of each word in a context window of size three around the medical term.

[1]Negative prefixes: ab, de, di, il, im, in, ir, re, un, no, mel, mal, mis. In retrospect, some of these are too general and should be tightened up in the future.

- The cluster identifier of head word
- The cluster identifiers of two preceding and two following verbs derived for lexico-syntactic features

As described above, considering clinical text properties, several features were newly designed with the minority classes in mind. In the next section, I analyze the contributions of each feature type in detail.

4.3 Evaluation of Assertion Classification Model

The input for this task is clinical texts and concept annotations. For gold annotations, the assertion of each problem concept was manually labeled by medical experts. Figure 4.2 illustrates a sample text with concepts and assertions.

The assertion annotation file contains medical problems with their assertions, one medical assertion per line. The assertion annotation format specified in the 2010 i2b2 Challenge Annotation File Formatting [1] is as follows:

`c="concept text" offset||t="concept type"||a="assertion value"`

where

`c`, `offset`, and `t` are defined as in Section 3.4.

`a` represents the assertion of the concept.

For example, in the sentence "*No pleural effusion or pneumothorax*," two medical problems, "*pleural effusion*" and "*pneumothorax*," have *absent* assertions.

4.3.1 Assertion Data Set

I evaluated performance on the assertion data from the 2010 i2b2 Challenge test set. The training set includes 349 clinical notes, with 11,967 assertions of medical problems. The test set includes 477 texts with 18,550 assertions. Table 4.1 shows the distribution of each assertion type in the training and test data. Two of the assertion categories (*present* and *absent*) accounted for nearly 90% of the instances in the data set, while the other four classes were relatively infrequent.

I do not attempt weakly-supervised learning for assertion classification as regarded as beyond the scope of this dissertation research. However, considering the imbalanced dataset,

```
1     She had a liver function test and amylase and lipase postoperatively
.     and she had a digoxin level of 1.0 on 06/04/05 .
2     The patient had a CBC on admission of 14.1 with a hematocrit of 33.8 .
3     Her CBC remained stable on 06/05/05 .
4     She had a white blood cell of 7.7 , hematocrit of 30.6 .
5     The patient had a MRSA nasal culture obtained on 06/03/05 , which
.     revealed rare staphylococcus aureus .
6     The patient had a chest x-ray on admission , which was clear .
7     No pleural effusion or pneumothorax .
```

(a) A Sample Text (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc2.txt)

```
c="a liver function test" 1:2 1:5||t="test"
c="amylase" 1:7 1:7||t="test"
c="lipase" 1:9 1:9||t="test"
c="a digoxin level" 1:14 1:16||t="test"
c="a cbc" 2:3 2:4||t="test"
c="a hematocrit" 2:10 2:11||t="test"
c="her cbc" 3:0 3:1||t="test"
c="a white blood cell" 4:2 4:5||t="test"
c="hematocrit" 4:9 4:9||t="test"
c="a mrsa nasal culture" 5:3 5:6||t="test"
c="rare staphylococcus aureus" 5:13 5:15||t="problem"
c="a chest x-ray" 6:3 6:5||t="test"
c="pleural effusion" 7:1 7:2||t="problem"
c="pneumothorax" 7:4 7:4||t="problem"
```

(b) A Sample Concept Annotation (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc2.con)

```
c="rare staphylococcus aureus" 5:13 5:15||t="problem"||a="present"
c="pleural effusion" 7:1 7:2||t="problem"||a="absent"
c="pneumothorax" 7:4 7:4||t="problem"||a="absent"
```

(c) A Sample Assertion Annotation (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc2.ast)

Figure 4.2: A Sample Text With Concept and Assertion Annotations

Table 4.1: Assertion Types Distribution

Assertion type	**Training**		**Test**	
	Count	**Percent**	**Count**	**Percent**
Present	8,051	67.3	13,025	70.2
Absent	2,535	21.2	3,609	19.5
Possible	535	4.5	883	4.8
Conditional	103	0.9	171	0.9
Hypothetical	651	5.4	717	3.9
Not patient	92	0.8	145	0.8
All	11,967	100.0	18,550	100.0

the weakly-supervised learning approach, which I apply to medical relation classification in Chapter 5, might be beneficial to this assertion task. Extending the assertion classification model to weakly supervised learning will be discussed more in Chapter 6.

4.3.2 Evaluation Metrics

I used three metrics to evaluate assertion classification: recall, precision, and F_1 score, a harmonic mean of recall and precision (giving equal weight to each) [266]. Each metric was microaveraged or macroaveraged across each assertion in the test set. Microaveraged metrics are computed by globally counting the total true positives, false negatives and false positives, whereas macroaveraged metrics are locally computed for each assertion class and then averages them. In other words, macroaveraged metrics give equal weight to each assertion class, whereas microaveraged metrics give equal weight to each assertion instance. Given the extremely unbalanced distribution of assertion types in the data set, I provided the results of both microaveraged that favors the dominant classes and macroaveraged that can be more appropriate for unbalanced class distribution. I used the official i2b2 Challenge evaluation script to calculate microaveraged measures. For macroaveraged measures, a new script was created to obtain average values for each assertion type.

4.3.3 Results for Assertion Classification

I have conducted a set of experiments to evaluate the performance of SVM-based classifiers. In the next subsection, I present the classification results when the classifier was trained with the full set of features described above. Then, I show how each feature class contributes to assertion classification.

The supervised learning system trained with the i2b2 training data showed quite good performance, with 94.5% microaveraged F_1 score. The best performing system in the 2010 i2b2 Challenge achieved an F_1 score of 93.6%. Please note that each score of microaveraged metrics is identical to each other, that is, Recall = Precision, because the counts of system predictions are always corresponding to the number of assertions in the reference standard. The macroaveraged F_1 score was 81.4%, much lower because of weak recall rates of some minority classes, especially 26.9% recall for *conditional*. Table 4.2 shows the results produced with the supervised classifier. The assertion classifier reached over 95% F_1 score for two dominant classes: *present* and *absent*. For other minority classes including *not patient*

Table 4.2: Results Produced With the Supervised Assertion Classifier

Assertion type	**Recall**	**Precision**	**F_1 score**
Present	98.0	95.1	96.5
Absent	95.6	95.7	95.6
Possible	59.3	81.2	68.6
Conditional	26.9	78.0	40.0
Hypothetical	87.9	91.0	89.4
Not patient	80.0	95.9	87.2
Macroavg	74.6	89.5	81.4
Microavg	94.5	94.5	94.5

(named for *Not associated with the patient*), the classifier obtained high precision, mostly over 80%, with variable recall.

4.3.4 Analysis and Discussion

I use the prediction confusion matrix to explain the performance of the assertion classifier for each assertion type on the test set. Table 4.3 displays counts of true positives (bolded), false positives, and false negatives of each category in a confusion matrix. Most of the *conditional* assertions were frequently misclassified as *present* with 114 false negatives and 13 false positives. For *possible*, the classifier produced 325 false negatives predicted as *present*. Several *not patient* assertions were misclassified as *present* or *absent* assertions.

I performed ablation tests for the assertion classifiers to measure the contribution of each of the five subsets of features explained above. An ablated classifier was tested by excluding the feature set specified in each row header in Table 4.4. The columns named "Impact" in Table 4.4 shows the F_1 score difference between the ablated classifier and the complete system.

As shown in Table 4.4, adding any feature set resulted in better performance of the assertion classifier. Removing the lexical features showed the sharpest decline in both macroaveraged and microaveraged F_1 scores. More specifically, the lexical features increased macroaveraged and microaveraged F_1 scores by 4.2 ($= 81.37\% - 77.17\%$) and 1.5 ($= 94.50\% - 93.05\%$) respectively. The macroaveraged F_1 score dropped much more than microaveraged F_1 score when contextual or lexical features were removed. Apparently, this indicates they were more beneficial for minority classes while other features had less impact on the minority classes. Removing word embedding features led to the macroaveraged F_1 of 80.7% and the microaveraged F_1 of 94.3%.

Table 4.5 shows the detailed results of each ablated classifier for each feature type. The columns named "Rec," "Pre," and "F_1" in Table 4.5 present the recall, precision, and F_1 scores, respectively, obtained by each ablated classifier. The contextual features contributed to the performance on all assertion types. Removing the contextual features had the biggest impact on *not patient* type. The recall of *not patient* decreased from 87.2% to 62.8% without the contextual features. The contextual features helped also detect more *conditional* cases. Allergy-related section headers, for example, "*Allergies*," "*Allergies and Medicine*

Table 4.3: Confusion Matrix of Assertion Predictions

	Classified as					
Gold	Present	Absent	Possible	Conditional	Hypothetical	Not patient
Present	**12,765**	119	89	13	39	0
Absent	144	**3,449**	13	0	1	2
Possible	325	16	**524**	0	18	0
Conditional	114	5	2	**46**	4	0
Hypothetical	63	4	17	0	**630**	3
Not patient	16	13	0	0	0	**116**

True positives (the diagonal elements) are bolded.

Table 4.4: Features Contribution to Assertion Classification

Feature	**Macroaveraged**		**Microaveraged**	
	F_1 score	**Impact**	**F_1 score**	**Impact**
- Contextual	78.1	-3.3	94.1	-0.5
- Lexical	77.2	-4.2	93.1	-1.5
- Lexico-syntactic	81.0	-0.3	94.3	-0.2
- Syntactic	80.6	-0.7	94.3	-0.2
- Word embedding	80.7	-0.7	94.3	-0.2

Table 4.5: Features Contribution for Each Assertion Type

Feature	Rec	Impact	Pre	Impact	F_1	Impact
All						
Present	98.0		95.1		96.5	
Absent	95.6		95.7		95.6	
Possible	59.3		81.2		68.6	
Conditional	26.9		78.0		40.0	
Hypothetical	87.9		91.0		89.4	
Not patient	80.0		95.9		87.2	
- Contextual						
Present	97.8	-0.2	94.8	-0.3	96.3	-0.2
Absent	94.9	-0.6	95.2	-0.4	95.1	-0.5
Possible	59.3	0.0	79.0	-2.2	67.8	-0.8
Conditional	24.6	-2.3	70.0	-8.0	36.4	-3.6
Hypothetical	86.9	-1.0	91.0	-0.1	88.9	-0.6
Not patient	62.8	-17.2	90.1	-5.8	74.0	-13.2
- Lexical						
Present	97.5	-0.5	93.7	-1.4	95.6	-1.0
Absent	93.4	-2.2	93.8	-1.9	93.6	-2.0
Possible	52.2	-7.1	79.4	-1.9	63.0	-5.6
Conditional	6.4	-20.5	78.6	0.6	11.9	-28.1
Hypothetical	84.5	-3.4	88.5	-2.6	86.5	-3.0
Not patient	80.0	0.0	91.3	-4.5	85.3	-1.9
- Lexico-syntactic						
Present	98.0	0.0	94.9	-0.2	96.4	-0.1
Absent	95.5	0.0	95.7	0.0	95.6	0.0
Possible	56.7	-2.6	79.3	-2.0	66.1	-2.5
Conditional	26.3	-0.6	79.0	1.0	39.5	-0.5
Hypothetical	87.6	-0.3	91.4	0.4	89.5	0.0
Not patient	80.7	0.7	95.9	0.0	87.6	0.4
- Syntactic						
Present	98.0	0.0	94.8	-0.2	96.4	-0.1
Absent	95.2	-0.4	95.4	-0.2	95.3	-0.3
Possible	57.5	-1.8	80.3	-1.0	67.0	-1.6
Conditional	26.3	-0.6	73.8	-4.2	38.8	-1.2
Hypothetical	87.5	-0.4	91.7	0.6	89.5	0.1
Not patient	79.3	-0.7	95.8	0.0	86.8	-0.4
- Word embedding						
Present	97.9	-0.1	94.9	-0.2	96.4	-0.1
Absent	95.4	-0.2	95.4	-0.3	95.4	-0.2
Possible	58.0	-1.4	80.9	-0.4	67.6	-1.0
Conditional	25.7	-1.2	78.6	0.6	38.8	-1.2
Hypothetical	87.6	-0.3	90.5	-0.6	89.0	-0.4
Not patient	76.6	-3.5	96.5	0.6	85.4	-1.8

Rec = Recall, Pre = Precision, and F_1 = F_1 score

Reactions," "*Allergies/Sensitivities*," "*Allergy*," and '*Medication Allergies*," were associated with conditional assertions. The precision of *conditional* decreased from 78.0% to 70.0% without the contextual features.

The lexical features also contributed to the performance on each assertion type. As discussed above, The F_1 score of each assertion type dropped the most when the lexical features were removed except *not patient*. The lexical features gave 28.1%, 5.6%, and 3.0% higher F_1 scores for *conditional*, *possible*, and *hypothetical* assertions respectively. The negative prefix features increased performance on the *absent* type.

The lexico-syntactic features did not help for *absent* and *hypothetical* assertions. The *possible* class benefitted the most from the lexico-syntactic features, with a 2.6% recall gain. I observed that many *possible* concepts were preceded by a question mark ('?') in the training corpus. Similar to the lexico-syntactic features, the syntactic features increased performance on the *possible* assertions. The syntactic features allowed more precise classification on *conditional* assertions, with a 4.2% precision gain.

The word embedding features helped more than lexico-syntactic and syntactic features for *not patient* assertions. These features provided different representation of words for more generalization to unseen words from the training corpus. Overall, each type of feature contributed to the performance of the assertion classifier. I also provide another table in Appendix F for the reader who would like to see the results grouped by assertion categories.

My assertion classifier compares favorably to three state-of-the-art systems. Table 4.6 shows the F_1 scores of other state-of-the-art systems for medical assertion classification. Two systems were created for participation in the 2010 i2b2/VA Challenge [46, 64] and one as postchallenge effort [16]. de Bruijn et al. [64] reached a macroaveraged F_1 score of 77.4% (the highest macroaveraged F_1 score reported in [64]). Clark et al. [46] produced a 80.2% macroaveraged F_1 score (not officially reported in [46] but computed by taking the numbers in the confusion matrix from [46]). My SVM-based classifier (the last row of Table 4.6) obtained a slightly higher F_1 score than other systems in both micro and macroaveraging.

4.4 Conclusions

I created an SVM-based assertion classifier that achieves state-of-the-art performance on assertion labeling for clinical texts. In this chapter, I described the features used for

Table 4.6: Comparison With Other State-of-the-Art Systems for Assertion Classification

System	**Macroaveraged**	**Microaveraged**
de Bruijn et al. (2011) [64]	77.4	93.6
Clark et al. (2011) [46]	80.2	93.4
Bejan et al. (2013) [16]	80.0	94.2
My SVM classifier	81.4	94.5

this task, presented the experimental results on the i2b2 test set, and investigated the improvements resulting from the addition of each type of features. My analysis showed that the contextual or lexical features contributed the most to the system's performance. Especially, the *conditional* assertion type benefitted the most from lexical features. However, performance on the minority classes still lags behind the dominant classes, so more work is needed in this area. I discuss potential future directions for research on assertion classification in Chapter 6.

The assertion classification model was needed for weakly supervised learning preparation of the medical relation classification task which I will discuss in the next chapter. I apply the trained model to classify the assertion of each medical problem concept extracted from unlabeled data.

CHAPTER 5

EXPLOITING UNLABELED TEXTS FOR MEDICAL RELATION CLASSIFICATION

Medical relation classification is the main topic of this chapter. Given a pair of medical concepts found in a sentence, a relation classification system must determine the type of relation that exists between the two concepts. My research focuses on the relation classification task introduced in 2010 for the i2b2 Challenge Shared Tasks [263]. This task involves recognizing eight types of relations between pairs of three types of medical concepts: problems, treatments, and tests. Note that this task aims to classify relations of given reference standard concepts.

A key challenge of this task is the extremely skewed class distribution across relation types. For example, five types of relations are defined between problems and treatments, but two of them (*None* and *TrAP* (treatment administered for problem)) account for 86% of the instances in the i2b2 Challenge data. Four relation types (*TrCP* (treatment causes problem), *TrIP* (treatment improves problem), *TrWP* (treatment worsens problem), and *TrNAP* (treatment not administered because of problem)) are distributed across the remaining 14% of the data. Each of these "minority" relations appears in just 2-6% of the data. Identifying these minority relations is extremely important from a practical perspective because they hold valuable information. For example, the dominant relations between problems and treatments are *TrAP* (administration of a treatment) and *None* (no relation at all). In contrast, the minority relations (*TrCP*, *TrIP*, *TrWP*, *TrNAP*) represent situations were a treatment *causes*, *improves*, *worsens*, or is *contraindicated* for a problem, which are arguably the most important types of situations to recognize.

Classifiers trained with supervised learning can perform relatively poorly on minority classes because there are few examples in labeled training data. Exploiting unlabeled data for training presents an opportunity to improve performance on these classes. Self-training that

solely depends on the confidence score would not guarantee the satisfactory classification of minority classes when most of the confident examples come from majority classes. I present two instance selection methods that can offer a more robust solution when the labeled data has a skewed class distribution and acquiring good quality minority class instances is difficult. The proposed instance selection methods are specifically aimed at improving performance on minority classes. To surmount the traditional self-training problem that primarily selects instances of dominant classes, these new methods can select a diverse and representative set of new instances from the unlabeled data.

The first instance selection method, called *Unlabeled Data Prototypes (UDP) Selection*, selects instances from clusters containing only unlabeled data. The most representative instance from each cluster is selected as additional training data. The second method, called *Labeled Data Counterparts (LDC) Selection*, selects instances from clusters containing both labeled and unlabeled instances. For each labeled instance, this method identifies its closest counterpart by selecting the unlabeled instance in the cluster that is most similar to it.

In this chapter, I introduce the types of medical relations that need to be classified and outline the distribution of labeled data. Then, I explain how I extract the medical concepts from unlabeled data and classify the assertions of medical concepts by application of my MCE system (Chapter 3) and assertion classification model (Chapter 4). Next, I present an SVM model which will be used as a baseline and a component in my weakly supervised framework. Then, I present my weakly supervised learning model. I explain how the examples of unlabeled data and labeled data are clustered together and elaborate on the two instance selection methods. Finally, I show the experimental results and compare the differences between these selection methods.

5.1 Labeled Data Description

First, I describe each type of relation that exists between two concepts. The examples provided for the relations are excerpted from the 2010 i2b2 Challenge relation annotation guidelines [4]. For each relation type, concepts involved in the relation type appear underlined.

- Medical problem—treatment (***Pr–Tr***) relations:

    1. Treatment *improves* medical problem (*TrIP*).

- *hypertension was controlled on hydrochlorothiazide*
- *infection resolved with antibiotic course*

2. Treatment *worsens* medical problem (*TrWP*).
   - *the tumor was growing despite the available chemotherapeutic regimen*
   - *culture taken from the lumbar drain showed Staphylococcus aureus resistant to Nafcillin*
3. Treatment *causes* medical problem (*TrCP*).
   - *Penicillin causes rash .*
   - *hypothyroidism following near total thyroidectomy*
4. Treatment is *administered* for medical problem (*TrAP*).
   - *He was given Lasix periodically to prevent him from going into CHF .*
   - *Dexamphetamine 2.5 mg. p.o. q. A.M. for depression*
5. Treatment is *not administered* because of medical problem (*TrNAP*).
   - *Relafen which is contraindicated because of ulcers .*
   - *Colace 100 milligrams po q day , hold for loose stools .*
6. Relation that does not fit into one of the above defined relationships (*NoneTrP*).

- Medical problem—test (***Pr–Te***) relations:

1. Test *reveals* medical problem (*TeRP*).
   - *an echocardiogram revealed a pericardial effusion*
   - *An abdominal ultrasound was performed showing no stones .*
2. Test *conducted to investigate* medical problem (*TeCP*).
   - *a VQ scan was performed to investigate pulmonary embolus*
   - *chest x-ray done to rule out pneumonia*
3. Relation that does not fit into one of the above defined relationships (*NoneTeP*).

- Medical problem—medical problem (***Pr–Pr***) relations:

1. Medical problem *indicates* medical problem (*PIP*).
   - *Azotemia presumed secondary to sepsis*
   - *a history of noninsulin dependent diabetes mellitus , now presenting with acute blurry vision on the left side .*
2. Relation that does not fit into PIP relationship (*NonePP*).

I used the i2b2/VA 2010 Shared Task corpus for this research, which consists of a training set of 349 annotated clinical notes and a test set of 477 annotated clinical notes. This test set contains 45,009 annotated medical concepts with 9,069 relations that occur in the same sentence. Table 5.1 shows the distribution of each relation type in the training and test data.

The test set contains 6,949 *Pr–Tr* pairs that occur in the same sentence, of which 3,463 are positive examples (participate in a relation) and 3,486 are negative examples (*NoneTrP*). *Pr–Te* relations include 3,620 positive and 2,452 negative examples (*NoneTeP*). *Pr–Pr* relations include 1,986 positive and 11,190 negative examples (*NonePP*). As seen in Figure 5.1, the class distributions across *Pr–Tr* and *Pr–Te* relation types are extremely skewed.

Among *Pr–Tr* relations, four "minority" classes, *TrCP*, *TrIP*, *TrWP*, *TrNAP*, are distributed across 14% of the data. Each of these relations appears in just ∼2–6% of the data. Among the *Pr–Te* relations, *TeCP* is the minority class, accounting for <10% of the instances. My goal is to improve relation classification with an emphasis on these minority classes by exploiting large amounts of unlabeled clinical texts. Since there is only one type of *Pr–Pr* relation (*PIP*), I focused exclusively on the *Pr–Tr* and *Pr–Te* relations in my efforts.

5.2 Data Preparation for Weakly Supervised Learning

For this dissertation research, I also used texts from the MIMIC II Clinical Database [220], which contains various types of clinical notes: discharge summaries, nursing progress notes, cardiac catheterization notes, ECG reports, radiology reports, and echocardiography reports. From this data set, I used 26,485 discharge summaries after filtering out notes with insufficient text content (<500 Bytes).

For weakly supervised learning preparation, I had to identify the medical concepts in the unlabeled data and classify the assertion of each medical problem concept. For concept

Table 5.1: Relation Types Distribution

Relation type	Training		Test	
	Count	Percent	Count	Percent
Pr–Tr				
TrIP	107	2.5	198	2.8
TrWP	56	1.3	143	2.1
TrCP	296	6.9	444	6.4
TrAP	1422	32.9	2487	35.8
TrNAP	106	2.5	191	2.7
NoneTrP	2329	54.0	3486	50.2
Pr–Te				
TeRP	1733	48.6	3032	49.9
TeCP	303	8.5	588	9.7
NoneTeP	1533	43.0	2452	40.4
Pr–Pr				
PIP	1239	14.4	1986	15.1
NonePP	7349	85.6	11190	84.9

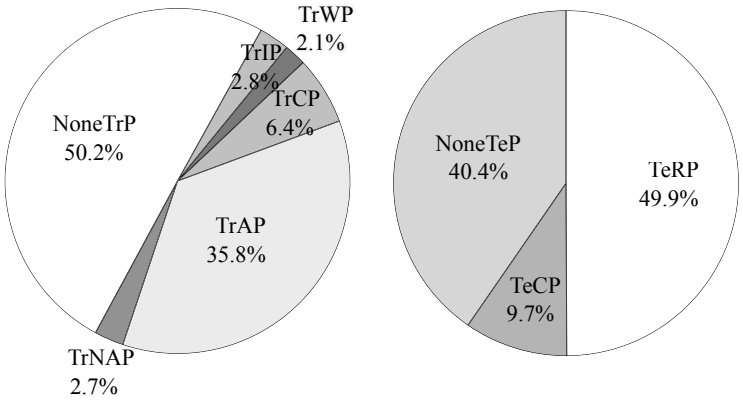


Figure 5.1: Distribution of Treatment and Test Relation Types in the Test Set

extraction, I used the stacked learning ensemble described in Chapter 3. From the texts, 4,108,054 medical concepts were extracted with 1,121,405 treatments, 1,306,556 tests, and 1,680,093 medical problems. For assertion classification of medical problems, I used the assertion classifier described in Chapter 4. The assertion classifier was trained with the i2b2 training data. The classifier assigned to each medical *problem* concept extracted from the the unlabeled data one of six labels: *present*, *absent*, *hypothetical*, *possible*, *conditional*, or *not associated with the patient*. Using the predicted concepts assigned to the unlabeled data, I created a large set of relation pairs to generate feature vectors for weakly supervised learning and clustering.

I used CLUTO [123], a data clustering software that has been widely used in various tasks, to create clusters containing both labeled (i2b2 training) and unlabeled data: 517,689 pairs of *Problem* and *Treatment* concepts and 455,272 pairs of *Problem* and *Test* concepts. The same feature vectors generated for SVM classification (to be discussed in Section 5.3) were reused with the clustering algorithm. The similarity between two instances was computed as the cosine between the instance vectors. Given two instance vectors, A and B, the cosine similarity is calculated as follows:

$$cos(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}} \tag{5.1}$$

To determine the number of clusters, I use the root-mean-square standard deviation (RMSSD). *RMSSD* is a measure of homogeneity within clusters and large RMSSD values indicate that clusters are not homogeneous [156]. Given a set of clusters, $C = \{c_1, c_2, ..., c_n\}$, *RMSSD* is defined as follows:

$$\begin{aligned} RMSSD(C) &= \sqrt{\frac{\sum_{i=1}^{n} \sum_{x \in c_i} \|x - m_i\|^2}{\sum_{i=1}^{n} (t_i - 1)}}, \\ \text{where } c_i &= \text{the } i_{th} \text{ cluster}, \\ m_i &= \text{center of } c_i, \\ t_i &= \text{number of instances in } c_i, \\ \|X\|^2 &= X^\top X \end{aligned} \tag{5.2}$$

I ran a series of clustering processes with different numbers of clusters, K, and calculated the RMSSD for each K. I tried 20 different cluster sizes aimed at having the average number of members per cluster range from 40 to 800 (i.e., K = the number of instances $\times$ n, n = 1/800, 2/800, ..., 19/800, 20/800). When I set n to 1/800 and 20/800 (= 1/40), I expected that on average 800 and 40 members would exist in each cluster, respectively.

The curves of RMSSD are generally either upward or downward. I used the shift point of the curves to determine the appropriate number of clusters. For each of the *Pr–Tr* and *Pr–Te*, I then detected the shift point (also known as the "Knee" point) of its RMSSD curve based on the Satopää et al. [223] method. Figure 5.2 shows the RMSSD curve of *Pr–Tr* clusters. According to [223], the Knee points are the local minima when the RMSSD curve is rotated θ degrees counterclockwise about (x_{min}, y_{max}) through the line drawn by connecting the points (x_{min}, y_{max}) and (x_{max}, y_{min}). In Figure 5.2, the point (x_{min}, y_{max}) is (647, 0.837) and (x_{max}, y_{min}) is (12940, 0.748). Depending on the curve, the local minima can be more than one but only one minimum existed in both RMSSD curves of *Pr–Tr* and *Pr–Te* clusters.

The cluster sizes of 4,529 and 3,414 were identified as the Knee points for the *Pr–Tr* and *Pr–Te* relation clusters respectively. In the following paragraphs, I will describe my supervised classification models and then present the instance selection methods based on clustering unlabeled data.

5.3 Supervised Relation Classification

I created three supervised learning classifiers (one for each category of concept pairs: *Pr–Tr*, *Pr–Te*, and *Pr–Pr*) using a rich set of features. I applied the Stanford CoreNLP tool [157] for tokenization, lemmatization, POS tagging, and phrase chunking. The system architecture for supervised relation classification is depicted in Figure 5.3.

5.3.1 Feature Set Description

I trained SVM classifiers with a linear kernel using the LIBLINEAR software package [80]. The multiclass SVM classifiers use six types of features associated with a pair of concepts $<C_1, C_2>$:

- Assertion Features:

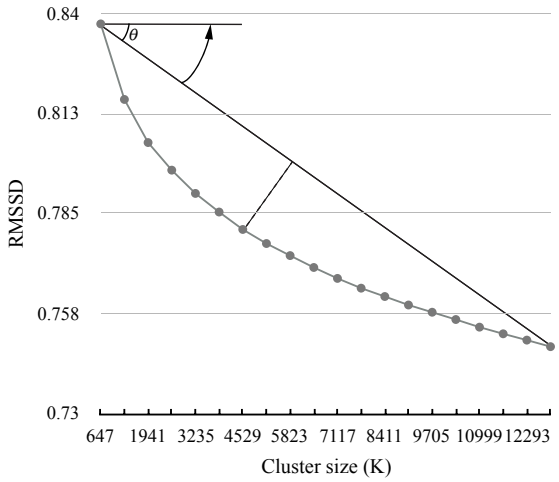


Figure 5.2: RMSSD Curve of *Pr–Tr* Clusters

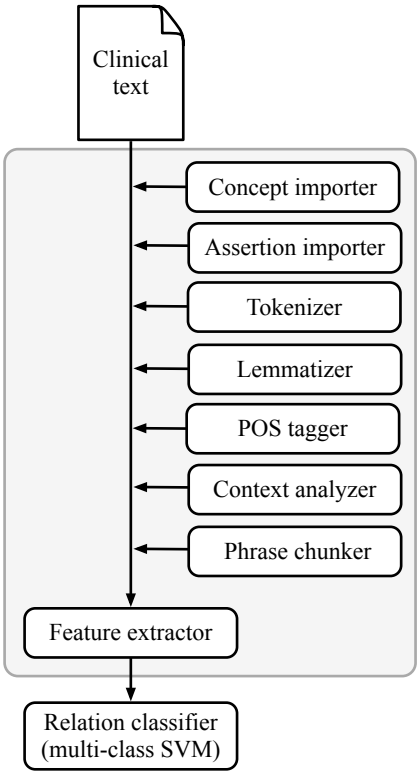


Figure 5.3: System Architecture for Supervised Relation Classification.

  - For each medical problem concept, I create a feature for the assigned assertion type.

- Context Features:

  - To compensate for the absence of assertions for treatment and test concepts, I incorporated the *ConText* algorithm [34] at the sentence level to detect three types of contextual properties for each concept: *negation*, *hypothetical*, and *historical*.

  - I also created a second set of *ConText* algorithm properties restricted to the six-word context window around C_1 and C_2 (three words on the left of C_1 and three words on the right of C_2).

- Distance Features: These features were designed to help the classifiers distinguish between concept pairs in close proximity that probably have a relation and distant pairs that probably have no relation between them.

  - I created features to represent the distance between concepts C_1 and C_2 by counting the number of words.

  - Features that specify whether the distance is the shortest or the farthest relative to other relations in the same sentence.

  - I created three features to measure the number of *treatment*, *test*, and *problem* concepts appearing before or after the pair.

  - Three features that specifiy whether any *Pr–Tr*, *Pr–Te*, and *Pr–Pr* relations exist between C_1 and C_2.

  - I created features that represent the sequence of chunk tags between C_1 and C_2.

- Lexical Features:

  - I created lexical features for the words contained in C_1 and C_2.

  - Bi-grams of words contained in C_1 and C_2.

  - The head words of C_1 and C_2

  - Two preceding and two following words for each of C_1 and C_2

  - The words between the two concepts
  - Also, I defined features for verbs that precede, follow, or occur between the concepts.

- Syntactic Features:
  - POS tags of two words on the right of C_1 and POS tags of two words on the left of C_2.

- Word Embedding Features:
  - In the same way as for assertion classification (Section 4.2), I used the cluster identifier of each word between the two concepts as a feature.
  - I also used the cosine similarity of the word embedding vectors for the heads of C_1 and C_2.

5.3.2 Training SVM Models

I randomly selected 149 ($= 349 - 200$, about 40% of the training set) documents from the training set as held-out data. I tuned LIBLINEAR's parameters to maximize the microaveraged F_1 score with the held-out data. After experimenting with different values on the development data, I set the cost parameter c to 0.06 for *Pr–Tr*, and 0.02 for *Pr–Te* and *Pr–Pr*. Also, the weights of negative examples were set to 0.2 for *Pr–Tr* and *Pr–Te* and 0.3 for *Pr–Pr*. The lower the weight for instances with no relation, the higher recall was obtained on held-out data.

Although the classifiers showed good performance under the microaveraged scoring metrics, performance on the minority classes was weak. As shown earlier, the class distributions are extremely skewed and the minority classes are relatively rare. To reduce the performance gap between the dominant classes and the minority classes, I also experimented with retraining the model by assigning higher weights to the minority classes to increase the importance of minority classes being classified correctly. It did not yield an increase in macroaveraged F_1 score and more detailed results will be reported in Section 5.5. To improve performance across the different relation classes, I extend my methods to weakly supervised learning described in the following paragraphs.

5.4 Exploiting Unlabeled Data for Relation Classification

To take advantage of the large amounts of unlabeled clinical notes that are available, I explored an iterative weakly supervised learning framework. I developed two novel methods for instance selection that are specifically aimed at improving performance on minority classes. My general framework involves the following steps:

1. A classifier is trained with supervised learning using the labeled training data.
2. The classifier is applied to the unlabeled data so that each unlabeled instance receives a predicted label.
3. A subset of the unlabeled instances is selected and then added to the set of labeled data (using the classifier's predictions as the labels).
4. The classifier is retrained using the (larger) set of labeled data.

This process repeats until a stopping criterion is met (e.g., for a fixed number of iterations or until no new instances can be labeled). Figure 5.4 shows the process for a learning mechanism for medical relation classification exploiting unlabeled data.

This paradigm is generally known as self-training, where the most common method for instance selection (step 3) sorts the instances based on the confidence scores produced by the classifier (i.e., confidence in the predicted labels) and then selects the most confidently labeled instances. This traditional self-training approach, however, tends to select instances of the dominant classes much more often than the minority classes because the classifier is more confident in its predictions for the dominant classes.

This issue motivated me to explore new methods for instance selection that try to create a diverse and representative set of new instances from the unlabeled data. Consequently, I developed two new methods for instance selection that first cluster the unlabeled data to identify groups of similar instances. Both methods generate clusters and assign labels to the instances in the same way.

First, the labeled and unlabeled instances are combined into a single dataset and the clustering algorithm (described in Section 5.2) is applied. Once the classifier predicts the label of each unlabeled instance, I consider the instances with a high confidence score as candidates for selection. In each iteration, I sort the instances based on the confidence scores

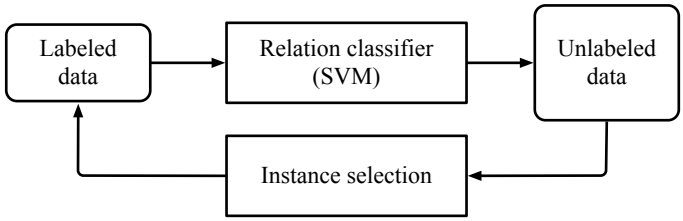


Figure 5.4: Learning Mechanism for Relation Classification.

produced by the classifier and an instance is a candidate for selection when it is ranked in the top 25% per class.

5.4.1 Unlabeled Data Prototypes Selection

The first instance selection method, called *Unlabeled Data Prototypes (UDP) Selection*, selects instances from clusters containing only unlabeled data.

1. I disregard clusters that contain labeled instance.

2. I compute the purity of each cluster and identify clusters where the highly confident cluster members have the same **positive** (participate in a relation) relation type (i.e., cluster purity = 1). Any instances that were already added to the set of labeled data are excluded from the calculation of cluster purity.

3. I discard clusters with purity < 1 because the instances are similar but the classifier's predictions are inconsistent, so the predictions are suspect.

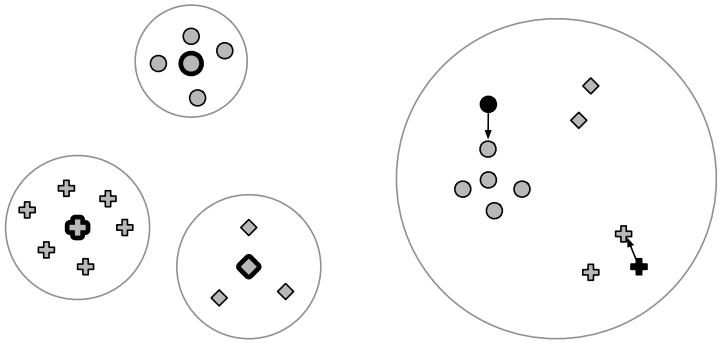
4. The most representative instance from each cluster is then selected as additional training data, based on average cosine similarity (defined in Equation 5.1) with other cluster members. An instance is excluded from the selection when it is exactly similar (i.e. cosine similarity = 1) to any instances that were already added to the set of labeled data.

The intuition behind this approach is twofold: (1) The instances in these clusters are different from the training instances, because no labeled instance was put in these clusters. Therefore, they could represent some new type of information found in the unlabeled data. (2) Choosing one representative instance from each cluster ensures that the set of selective instances will be diverse. This method is illustrated in Figure 5.5(a). Gray-colored instances represent unlabeled data.

5.4.2 Labeled Data Counterparts Selection

Assuming that unlabeled data will be similar to labeled data when they coexist in the same cluster, my second method, called *Labeled Data Counterparts (LDC) Selection*, selects instances from the clusters containing both labeled and unlabeled instances.

1. I disregard clusters that contain no labeled instance.



(a) *UDP* (b) *LDC*

Figure 5.5: Clustering-Based Instance Selection

2. For each instance labeled with a **positive** relation type in the original training data, the unlabeled instance most similar to it in the same cluster is selected. Similar to *UDP*, an instance is excluded from the selection when it is exactly similar to any instances in the labeled data.

My intuition is that this approach will acquire new training instances that share features with the original labeled data and maintain the similar class distribution. This method is illustrated in Figure 5.5(b). Black-colored instances represent labeled data and gray-colored instances represent unlabeled data. In the next sections, I compare the performance of self-training with confidence-based instance selection against my new *UDP* and *LDC* instance selection methods.

5.5 Evaluation of Relation Classification

The input for this task is clinical texts, concept annotations, and assertion annotations. Relations from the reference standard data are used to train the classifiers. Figure 5.6 illustrates a sample text with concepts, assertions, and relations.

The relation annotation file contains two medical concepts with their relation, one relation per line. The relation annotation format specified in the 2010 i2b2 Challenge Annotation File Formatting [1] is as follows:

`c="concept text" offset||r="relation type"||c="concept text" offset`

where

`c` and `offset` are defined as in Section 3.4.

`r` represents the type of relation the two concepts have.

The second `c` and `offset` represent the other concept in the relation.

5.5.1 Evaluation Metrics

I used three metrics to evaluate relation classification: recall, precision, and F_1 score. Each metric was microaveraged or macroaveraged across each relation in the test set. I used the official i2b2 Challenge evaluation script to calculate microaveraged measures. For macroaveraged measures, I created a new script to average the scores across relation types. The macroaveraged F_1 score is the harmonic mean of the macroaveraged recall and precision.

```
 1     PRINCIPAL DIAGNOSIS :
 2     Wolfe-Parkinson White Syndrome .
 3     ASSOCIATED DIAGNOSIS :
...    ...
 9     He has a history of chest pain and in January 1993 underwent a cardiac
 .     catheterization at Ph University Of Medical Center which revealed an
 .     occluded right coronary artery and a 40-50% proximal stenosis .
10     He subsequently had an echocardiogram in December 1994 which showed
 .     normal left ventricular size and systolic function .
```

(a) A Sample Text (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc1.txt)

```
c="wolfe-parkinson white syndrome" 2:0 2:2||t="problem"
...
c="chest pain" 9:5 9:6||t="problem"
c="a cardiac catheterization" 9:12 9:14||t="test"
c="an occluded right coronary artery" 9:23 9:27||t="problem"
c="a 40-50% proximal stenosis" 9:29 9:32||t="problem"
c="an echocardiogram" 10:3 10:4||t="test"
```

(b) A Sample Concept Annotation (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc1.con)

```
c="wolfe-parkinson white syndrome" 2:0 2:2||t="problem"||a="present"
...
c="chest pain" 9:5 9:6||t="problem"||a="present"
c="an occluded right coronary artery" 9:23 9:27||t="problem"||a="present"
c="a 40-50% proximal stenosis" 9:29 9:32||t="problem"||a="present"
```

(c) A Sample Assertion Annotation (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc1.ast)

```
c="a cardiac catheterization" 9:12 9:14||r="TeCP"||c="chest pain" 9:5 9:6
c="a cardiac catheterization" 9:12 9:14||r="TeRP"||c="an occluded right
coronary artery" 9:23 9:27
c="a cardiac catheterization" 9:12 9:14||r="TeRP"||c="a 40-50% proximal
stenosis" 9:29 9:32
```

(d) A Sample Relation Annotation (Excerpted from https://www.i2b2.org/NLP/Relations/assets/doc1.rel)

Figure 5.6: A Sample Text With Concept, Assertion, and Relation Annotations

5.5.2 Statistical Significance Testing

I used paired t test to measure statistical significance [283]. The null hypothesis assumes that the two methods are not significantly different. The alternative hypothesis assumes that there is a significant difference between the two methods. The significance level, α, was set at 5%.

As recommended by [283], I ran 1,048,576 trials to calculate the statistical significance between two methods for each metric. For each trial:

1. The outputs of the two methods for each test instance are paired up.
2. Randomly swap the two output, based on a virtual "coin toss" (50/50 chance of swapping).
3. Check how the shuffle produces a difference in the metric of interest (e.g., recall, precision, or F_1 score).

Finally, compare the two sets of results and calculate the statistical significance for the metric of interest.

I have conducted an extensive set of experiments to evaluate the performance of supervised classifiers and weakly supervised learning with different instance selection methods. I evaluated performance on the relation data from the 2010 i2b2 Challenge test set. In the next subsection, I present the classification results of supervised classifiers and compare them with weakly supervised learning results.

5.5.3 Supervised Learning Results

Table 5.2 shows the results produced with the supervised classifiers, which were trained to optimize for microaveraged measures. This baseline supervised learning system was trained with the i2b2 training data and achieved microaveraged scores of 74.9% recall, 73.7% precision, and 74.3% F_1 score.

As the state-of-the-art in medical relation classification, Zhu et al. [302] research produced a 74.2% microaveraged F_1 score on the 2010 i2b2 Challenge dataset. Although the supervised classifiers achieve overall performance comparable to state-of-the-art relation classification systems, performance on the minority classes lags far behind the dominant classes. The F_1 score of *TrWP* was only 7.6% with a recall of 4.2%. Most of the *TrWP* instances were

Table 5.2: Results Produced With the Supervised Relation Classifiers

Relation type	**Recall**	**Precision**	**F_1 score**
ALL	74.9	73.7	74.3
Treatment Avg.	67.4	68.9	68.2
TrIP	31.8	63.6	42.4
TrWP	4.2	42.9	7.6
TrCP	52.3	59.5	55.6
TrAP	79.9	71.2	75.3
TrNAP	25.1	49.5	33.3
Test Avg.	82.9	81.5	82.2
TeRP	90.3	82.7	86.3
TeCP	45.1	71.4	55.3
PIP	73.2	67.9	70.4

misclassified because of the very low prevalence of their cases. For example, consider a *TrWP* case from the test set, "*She has a known diagnosis of myelodysplasia that has become recalcitrant to Procrit*". The medical problem *myelodysplasia*, the treatment *Procrit*, and possibly a keyword *recalcitrant* never appeared in the training data. The lower performance on minority classes is more apparent when focusing on macroaveraged scores. Based on macroaveraging, this system reached 50.2% recall, 63.6% precision, and 56.1% F_1 score.

I also experimented with decreasing the weights of negative examples to help increase recall on minority classes. This did not yield an increase in macroaveraged F_1 score because of a substantial drop in precision. For instance, when I set the weight of negative examples to 0.04 (decreasing the weight by five from 0.2) for both *Pr–Tr* and *Pr–Te* relations, the system reached 55.4% macroaveraged recall, 54.5% precision, and 55.0% F_1 score. Adjusting the importance of different relation types by assigning different weights also did not affect performance very much. When I increased the weights of *TrIP*, *TrWP*, and *TrNAP* by one hundred to 1, and 30 times for *TrCP* to 1 considering the distribution of relation types in the labeled data, the classifiers achieved a recall gain of 1.8 (from 50.2% to 52.0%), but with a corresponding precision loss of 1.9 (from 63.6%to 61.7%) in macroaveraging.

5.5.4 Comparing Supervised and Weakly Supervised Learning Results

I evaluated the performance of self-training with traditional confidence-based instance selection (called **Self-training** below), and instance selection with my new *UDP* and *LDC* methods. I ran all of the weakly supervised learning methods for 20 iterations.[1]The number of iterations was determined after experimenting on held-out data.

For self-training, I only selected positive examples (pairs of concepts with relations) from the unlabeled data to augment the labeled data. For each iteration, I added K newly labeled examples, where K = the number of positive examples in the original training data. My intention was to be conservative in adding new examples with predicted labels to maintain the importance of the original labeled data. To maintain the same class distribution, I imposed that the number of newly labeled examples for each positive class should not exceed the number of examples in the original training data.

---

[1]The number of iterations was determined after experimenting on held-out data. The overall macroaveraged F_1 score showed a downward trend after 20 iterations.

Table 5.3 shows results for each class and macroaveraged F_1 scores for the *Pr–Tr* and *Pr–Te* relations. For each relation type, the best results appear in boldface. Results that are significantly different from the supervised learning results at the 95% significance level are preceded by an asterisk (*). Table 5.4 shows the macroaveraged overall recall, precision, and F_1 score of each method.

Self-training with confidence-based instance selection produced the best F_1 score on the *TrCP* and *TrNAP* classes. For *TrWP* and *TeCP*, self-training's performance was significantly different than supervised learning. Self-training yielded the overall average recall, precision and F_1 score of 54.5%, 60.0% and 57.1%, respectively (second row in Table 5.4).

Both the *UDP* and *LDC* instance selection methods produced higher macroaveraged F_1 scores than self-training. The *UDP* method (third column of Table 5.3) produced the best F_1 score of 49.3% on the *TrIP* class. The F_1 scores for *TrIP* and *TeCP* were significantly higher than for supervised learning. The *LDC* method (last column of Table 5.3) produced the highest F_1 scores on most of the relation classes. It obtained the best macroaveraged F_1 scores for Treatment and Test. For *TrIP*, *TrWP*, *TeRP*, and *TeCP*, the performance of *LDC* method was significantly better than supervised learning. The *LDC* method also produced improvements on the majority classes (*TrAP* and *TeRP*). There was no sacrifice on the majority classes.

Finally, I tried to combine the *UDP* and *LDC* methods. New instances were selected separately by the *UDP* and *LDC* methods and then the combined set of instances was added to the labeled data. However, this system produced an F_1 score of 58.0% (last row in Table 5.4), so did not outperform the *LDC* method on its own.

Figure 5.7 shows the macroaveraged F_1 scores per epoch. The gap between the supervised learning (the straight line at 56.1%) and each instance selection method indicates the performance difference between them. The solid gray line with circle dots represents the *LDC* method. The *LDC* produced higher F_1 scores than *UDP* or self-training methods in each iteration. The *LDC* method produced the best overall macroaveraged F_1 score of 58.5%. The dashed gray line with triangle dots represents the combination of *UDP* and *LDC* methods. It got slightly lower F_1 scores than *LDC* in most epochs. The black dashed line with diamond dots stands for the *UDP* method and the black solid line with square dots represents self-training method. The *UDP* method reached 57.9% macroaveraged F_1

Table 5.3: F_1 Score of Each Method on the Test Set

Relation type	Supervised	Self-training	UDP	LDC
Treatment Avg.	46.2	48.0	48.9	**49.7**
TrIP	42.4	46.0	***49.3**	*47.4
TrWP	7.6	*16.3	12.3	***19.2**
TrCP	55.6	**56.8**	55.5	53.1
TrAP	75.3	75.4	**75.8**	**75.8**
TrNAP	33.3	**35.4**	33.1	33.6
Test Avg.	72.0	72.6	72.8	**73.1**
TeRP	86.3	86.3	86.3	***86.7**
TeCP	55.3	*58.5	*59.2	***59.5**

Table 5.4: Overall Macroaveraged Scores for Each Method on the Test Set

Method	Recall	Precision	F_1 score
Supervised	50.2	63.6	56.1
Self-training	54.5	60.0	57.1
UDP	55.0	61.1	57.9
LDC	54.9	62.5	58.5
UDP+LDC	54.9	61.4	58.0

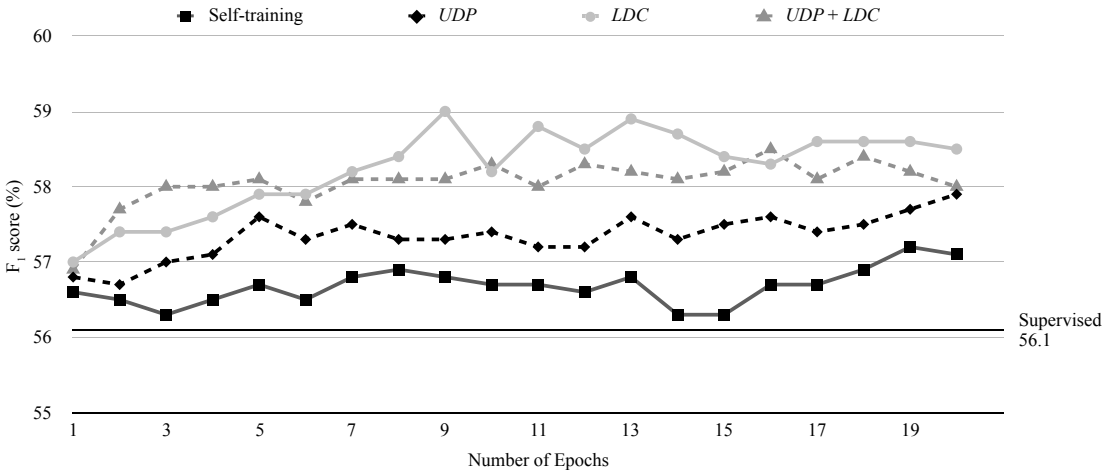


Figure 5.7: Macroaveraged F_1 Score of Each Method per Epoch

score. Self-training did not outperform other instance selection methods during iterations. It produced a macroaveraged F_1 score of 57.1%.

Table 5.5 shows the performance of other state-of-the-art systems for medical relation classification compared to the results from the *LDC* method. Two systems were created for participation in the 2010 i2b2/VA Challenge [64, 213] and one was subsequent research [302] to [64]. Macroaveraged scores were not reported in [302]. My *LDC* method (the last row of Table 5.5) obtained a higher macroaveraged F_1 score than other systems with a comparable microaveraged score.

5.5.5 Timing Analysis

I calculated the times for training supervised classification models (1st row in Table 5.6), predicting medical relations with the supervised classifiers (2nd row in Table 5.6), clustering labeled and unlabeled instances (3rd row in Table 5.6), and selecting a subset of the unlabeled instances by the *UDP* or *LDC* methods (4th and 5th rows in Table 5.6, respectively). All measurements were performed on a MacBook Pro with a 2.5 GHz Intel Core i7 processor and 16 GB of memory. The training set, test set, and unlabeled data include 349, 477, and 26,485 text files, respectively. Clusters contained 517,689 pairs of problem and treatment concepts and 455,272 pairs of problem and test concepts.

Training supervised classification models took 1.1 seconds. The supervised classifiers spent about 0.4 seconds predicting the relations on the i2b2 test set. Clustering took about 28 minutes but it was needed only once before the iterative instance selection process (*UDP* or *LDC*) started. Both the *UDP* and *LDC* methods selected new training instances within 3 minutes in each iteration. About one hour was needed for 20 iterations of both methods to apply the relation classifiers to the unlabeled data, select new instances, and retrain the classifiers.

5.5.6 Analysis and Discussion

I performed ablation testing of the supervised learning system to evaluate the impact of each feature set based on microaveraged and macroaveraged scores, separately. If some features have more impact for macroaveraged scores than microaveraged scores, then my hypothesis is that they are especially important features for minority classes. The row header in Table 5.7 specifies the feature set that has been ablated. The columns named "Impact"

Table 5.5: F_1 Scores of Other State-of-the-Art Systems for Relation Classification

System	Macroaveraged	Microaveraged
Rink et al. (2011) [213]	53.7	73.7
de Bruijn et al. (2011) [64]	51.2	73.1
Zhu et al. (2013) [302]	N/A	74.2
LDC	58.5	74.3

Table 5.6: Task Time Measurement

Model	Time
Supervised Classifier (Training)	1.1s
Supervised Classifier (Test)	0.4s
Clustering	28m
UDP	2m 42s
LDC	2m 44s

m = Minute and s = Second

Table 5.7: Features Contribution to Relation Classification

Feature	Macroaveraged		Microaveraged	
	F_1 score	Impact	F_1 score	Impact
All	56.1		74.3	
- Assertion	55.4	-0.7	73.8	-0.5
- Contextual	55.4	-0.7	74.2	-0.1
- Distance	55.2	-0.9	72.4	-1.9
- Lexical	49.0	-7.1	70.1	-4.2
- Syntactic	56.6	0.5	74.1	-0.2
- Word embedding	55.8	-0.3	73.8	-0.5

in Table 5.7 present the F_1 score difference between the ablated classifier and the complete system.

Every feature set contributed to the performance of the supervised classifiers except that syntactic features did not increase macroaveraged F_1 score. The macroaveraged F_1 score dropped the most when the lexical features were removed. This suggests that exploiting unlabeled data could be especially beneficial for the minority classes by bringing in new lexical features. The F_1 scores of *TrIP*, *TrNAP*, and *TeCP* decreased from 42.4%, 33.3%, and 55.3% to 29.4%, 21.9%, and 42.4% respectively without the lexical features.

Next, I want to see how my self-training design that only selects positive examples and maintains the original class distribution compares to self-training that can add negative examples and does not consider the class distribution for newly added instances. So I created a self-training system with the latter design for comparison. After the first iteration, I computed the class distribution of the newly added instances. I realized that no instances of any minority class was found among the instances of either *Pr–Tr* or *Pr–Te* relations. Furthermore, *TrA*P instances only took up 0.7% of the selected instances. The remaining 99.3% of the data was all negative examples. For *Pr–Te* relations, I found 41% to be *TeRP* instances but no *TeCP* instances.

I also carried out an empirical analysis of self-training with confidence-based instance selection (my self-training design) to better understand its limitations. After clustering the unlabeled data, I counted the number of instances selected from each cluster during the first iteration. I found that most instances were selected from a small subset of the clusters: about 10% of the clusters provided over 78% of the newly selected unlabeled instances. This shows that selecting instances based only on confidence scores tends to yield a relatively homogenous set of new instances that is low in diversity. In other words, although self-training was able to keep the class distribution, the diversity of new instances for each class was low.

The two methods proposed in this dissertation research showed potential for overcoming these self-training limitations. The representative examples identified as prototype instances by *UDP* would be dissimilar to each other because they occurred in different clusters. In *LDC*, the labeled instances play an essential role in maintaining the diversity of newly added instances because each labeled instance selects the most similar unlabeled instance.

When the labeled instances are distributed across many clusters, the *LDC* method can contribute diverse addition because newly labeled instances would have the same diversity as the original labeled instances. I examined how the labeled instances are scattered around the clusters. 43.9% of clusters contained one or more labeled instances for *Pr–Tr* and 48.9% clusters for *Pr–Te*. They contained 2.2 labeled instances on average for *Pr–Tr* and 2.1 on average for *Pr–Te*. These clusters were used for the *LDC* method while the others (56.1% for *Pr–Tr* and 51.1% for *Pr–Te*) were considered for the *UDP* method.

Figure 5.8 shows the number of unlabeled instances that were added by each instance selection method during the first iteration. The white column represents the self-training method. In self-training, no examples of *NoneTrP* or *NoneTeP* were selected. For the other relation types, the same number of examples as the original training data were added to maintain the class distribution. The gray column represents the *UDP* method. In *UDP*, similar to self-training, no examples of *NoneTrP* or *NoneTeP* were selected because an instance was allowed for selection only when it came from clusters where all members agreed to one positive relation type. The number of selected instances by the *UDP* method was less than that of other methods. This is because clusters with purity < 1 were excluded from the selection. The black column represents the *LDC* method. The *LDC* method allowed some negative examples to be added. An unlabeled instance can be selected when it is the most similar to any labeled instance, including negative instances. The number of selected instances by the *LDC* method was less than that of self-training because (1) *LDC* excluded unlabeled instances that were exactly similar to any labeled instance or (2) more than one labeled instances might pick the same unlabeled instance. Note that the *LDC* method outperformed self-training with fewer new instances.

Table 5.8 displays the Recall, Precision, and F_1 results of *LDC* instance selection along with the total counts of true positives (TP) and the number and percentage of true positive gains (compared to supervised learning) in the rightmost column. The numbers in parentheses in the Recall, Precision, and F_1 columns indicate the difference between the supervised classifier and the *LDC* method. Results significantly different from supervised learning at the 95% significance level are preceded by an asterisk (*).

Table 5.8 shows that most of the minority classes benefitted substantially from the *LDC* method. The largest percentage gain came for *TrWP* where *LDC* correctly identified 17

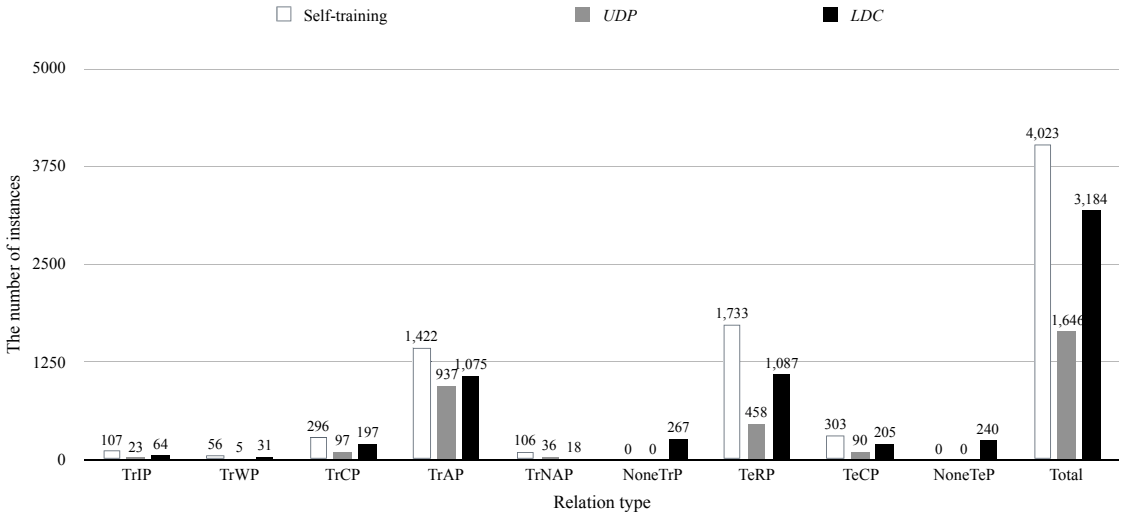


Figure 5.8: The Number of Unlabeled Instances Added During the First Iteration

Table 5.8: Results of LDC With Comparison to the Supervised Learning Model

Relation	Recall	Precision	F_1 score	TP	TP Gain (%)
Minority					
TrIP	*38.9 (+7.1)	60.6 (−3.0)	*47.4 (+5.0)	77	14 (+22.2)
TrWP	*11.9 (+7.7)	50 (+7.1)	*19.2 (+11.6)	17	11 (+183.3)
TrCP	*65.1 (+12.8)	*44.9 (−14.6)	53.1 (−2.5)	289	57 (+24.6)
TrNAP	23.6 (−1.6)	*58.4 (+9.0)	33.6 (+0.3)	45	-3 (−6.3)
TeCP	*57.7 (+12.6)	*61.4 (−10.0)	*59.5 (+4.2)	339	74 (+27.9)
Majority					
TrAP	*80.8 (+0.9)	71.3 (+0.2)	75.8 (+0.5)	2009	23 (+1.2)
TeRP	*88.5 (−1.8)	*85.0 (+2.4)	*86.7 (+0.4)	2682	-55 (−2.0)

TP = True positive

instances but the supervised learner only produced six true positives, resulting in a gain of 11 (183.3%). The majority classes also achieved slightly higher F_1 scores. The *LDC* method appears to be an effective way to improve recall on minority relation classes while maintaining good performance on the majority classes.

Table 5.9 displays counts of true positives (in the diagonal line), false positives, and false negatives for each *Pr–Tr* relation type produced by *LDC* in a confusion matrix. For the minority classes, the *LDC* method primarily produced false negatives that were predicted as *TrAP* or *NoneTrP*. For example, Most of the *TrWP* relations were misclassified as *TrAP* with 55 false negatives or as *NoneTrP* with 45 false negatives. Similarly, many *TrNAP* relations were misclassified as *TrAP* with 58 false negatives or *NoneTrP* relations with 50 false negatives.

5.6 Conclusion

I showed that clustering-based instance selection from unlabeled text data could improve performance on minority classes for relation type classification between medical concepts. Experimental results show that my clustering-based methods outperformed supervised classification, traditional self-training from unlabeled texts, and previous state-of-the-art systems based on macroaveraged scores. Importantly, overall microaveraged scores were also comparable, so these new instance selection methods maintain good performance on the majority classes. I believe that this approach offers a more robust solution for classification problems when the data has a highly skewed class distribution, acquiring manual annotations is expensive, but large quantities of unannotated text data are available.

Table 5.9: Confusion Matrix of *LDC* Method Predictions

	Classified as					
Gold	TrIP	TrWP	TrCP	TrAP	TrNAP	NoneTrP
TrIP	**77**	1	14	57	1	48
TrWP	7	**17**	18	55	1	45
TrCP	1	1	**289**	59	4	90
TrAP	22	7	91	**2009**	9	349
TrNAP	1	1	36	58	**45**	50
NoneTrP	19	7	196	578	17	**2669**

True positives (the diagonal elements) are bolded.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this chapter, I summarize my research on improving clinical information extraction from EHRs with multiple domain models and clustering-based instance selection. I report the findings and discuss contributions of my research. Then, I outline future work which can be considered.

6.1 Conclusions

In this section, I discuss claims, contributions, and key findings. This dissertation research demonstrates that combining IE models can improve medical concept extraction. It also shows that clustering-based instance selection methods can improve medical relation classification.

- *Claim 1: Ensemble methods with a combination of models trained on broad medical and specialty area texts can improve medical concept extraction on specialty notes.*

When a limited amount of annotated specialty area data is available, this research demonstrates that combining broad medical data and specialty area data can produce MCE models that achieve better performance on specialty notes than training with either type of data alone. I investigated the performance of MCE models on specialty notes (1) when trained on a broad medical corpus, (2) when trained on the same type of specialty data, and (3) when trained on a combination of models trained on both broad medical and specialty data. When training with a comparable amount of annotated data, I found that training with specialty texts outperformed training with broad medical texts. I achieved better performance for all three specialty areas by using a combination of both broad medical i2b2 data and specialty area models for training.

Another conclusion of this research is that a stacked ensemble with mixed domain models achieved good performance and offered some advantages over other approaches. I

investigated ensemble-based methods for medical concept extraction with a diverse set of information extraction models. The experimental results confirmed that ensemble architectures consistently outperformed individual IE models. The model trained on the union of the broad medical data and specialty data extracted more medical concepts. However, this model might not be the preferable solution when precision is more important than recall. For example, in medical relation classification, accurately extracted concepts are indispensable to select useful relation instances from abundant unlabeled data. My stacked learning ensemble produced higher precision by adjusting the influence of different component models.

The results also showed that the stacked learning ensemble has a significant practical advantage over the voting ensemble. The stacked learning ensemble was able to easily incorporate any set of individual concept extraction components because it automatically learned how to combine their predictions to achieve the best performance. Unlike previous research only employing the predictions and the confidence probabilities of individual models, my stacked generalization is firstly trained with a rich set of meta-features for more accurate medical concept extraction. Features dealing with the degree of agreement and consistency between the IE models were created to capture more concepts and detect their boundaries more precisely. This ensemble-based approach provided more flexible and robust integration of MCE models on specialty notes with a limited amount of labeled data.

As discussed in Chapter 2, the dominant research of MCE in the clinical domain has primarily been focused on broad medical texts. With relatively little research on MCE for specialized clinical texts, this research created new text collections that represent specialized areas of medicine: cardiology, neurology, and orthopedics. The new corpora to account for each specialty area were manually annotated by medical experts and used for training and evaluation of my approaches. For interested researchers who already have local Institutional Review Board approval to the BLUlab corpus, I am willing to cooperate with them by providing our reference annotations.

- *Claim 2: Clustering-based instance selection from unlabeled data can improve performance on minority classes in medical relation classification.*

To take advantage of the large amounts of unlabeled clinical notes that are available, I

explored an iterative weakly supervised learning framework. Because of extremely skewed class distributions, my supervised relation classifiers achieved high accuracy for the majority classes but performed poorly with the minority classes.

It motivated me to present new methods for instance selection: *Unlabeled Data Prototypes (UDP) Selection* and *Labeled Data Counterparts (LDC) Selection*. *UDP* selects instances from clusters containing only unlabeled data to represent some new type of information found in the unlabeled data. This method also ensures that the set of representative examples identified as prototype instances will be diverse. *LDC* selects instances from the clusters containing both labeled and unlabeled instances. This method can acquire new training instances that share features with the original labeled data and maintain the similar class distribution. These two methods are specifically aimed at improving performance on minority classes. They are based on clustering unlabeled data and can create a diverse and representative set of new instances from the unlabeled data. The experiment results showed that they achieved substantial performance gains for the minority relation classes compared to supervised learning and traditional self-training based on macroaveraged scores. These results also demonstrated that these methods maintained good performance on the majority classes. My *LDC* method obtained a higher macroaveraged F_1 score than other state-of-the-art systems with a comparable microaveraged score.

Benefiting from large amounts of unlabeled data with new instance selection methods based on similarity measures is a novel contribution of this dissertation. Two instance selection methods offer a more robust solution for classification problems when the data has a highly skewed class distribution and acquiring manual annotations is expensive, but large quantities of unannotated text data are available.

6.2 Future Work

This section presents some future work directions to extend the scope of this research. I discuss several ideas that might motivate further research.

6.2.1 Weakly Supervised Learning for Specialty Area Notes With Cross-Task Learning

One avenue for further study would be applying cross-task learning. Cross-task Learning [32, 62] is a method for simultaneously learning two different tasks using prior knowledge

that relates their outputs. For example, in syntactic chunking, an NNP (proper noun) can be extracted when the NNP is part of a named entity. In NER, a named entity can be extracted when it is a subsequence of a noun phrase.

A new cross-task framework can be explored to extract new concepts from unlabeled data more confidently and correctly than self-training. This cross-task learning can be formulated as simultaneously training models for medical concept extraction and medical relation classification. This learning method is similar to co-training requiring multiple learners. However, unlike co-training involving one specific task with different views of the data, this method can consider two different tasks simultaneously.

This cross-task learning requires two task components: a medical concept extractor and a medical relation classifier. Two independent components would provide different and complementary information to each other. The initial concept classifier and relation classifier are learned using labeled data. Then, the concept classifier extracts the candidates of medical concepts from unlabeled data.

Given a pair of medical concepts, with one known concept and the other possible concept (additional candidate concept), the relationship classifier is applied to determine whether the pair of medical concepts is related. The possible concept, out of relevant relations extracted by the relation classifier, is used to iteratively improve the models. The possible concepts can be produced by several strategies. The candidates can be 1) simply noun phrases and adjective phrases, 2) concepts tagged by knowledge-based systems, 3) the phrase with the lower confidence score than concept cutoff thresholds, or 4) combination of these methods.

6.2.2 Instance Selection for Assertion Classification

I created a supervised SVM model for assertion classification. Although the classifier achieved state-of-the-art performance, it performed relatively poorly on minority classes. Similar to relation classification, the class distribution is extremely skewed across assertion types. As seen in Chapter 4, two of the assertion categories (*present* and *absent*) account for nearly 90% of the instances in the data set. The other four classes are distributed across the remaining 10% of the data.

Considering the imbalanced dataset, the weakly-supervised learning approach can be beneficial to this assertion task. In medical relation classification, a pair of concepts that

participates in a relation is considered as a positive example. Conversely, a pair of concepts that are not related to each other is treated as a negative example. In assertion classification, a *present* assertion is one that does not fit into any other assertion category. *Present* assertions appear in about 70% of the data and would be collected as negative examples. The feature vectors generated for assertion classification can be reused for clustering. Then, my two instance selection methods (*UDP* and *LDC*) will be applicable to assertion classification.

6.2.3 Sentence-Level Selection Using Clustering

This research confirms that acquiring beneficial instances is important for performance improvement for weakly supervised learning. Instead of instances consisting of features associated with a pair of concepts, one can consider sentence-level selection to identify relevant unlabeled data.

For medical relation classification, I developed various types of features. The classifier with the set of features showed quite good performance. However, some information that is a good indicator of medical relations may often be unattainable as a feature. Sentence-level clustering can give more liberty; any information derived from sentences can be used for clustering. In my instance selection methods, each data point in the clusters represents a pair of concepts. On the other hand, in sentence-level clustering, each data point in the clusters represents a sentence.

For the *LDC* method, one can select sentences from the clusters containing sentences from both labeled and unlabeled data. For each sentence where any positive relation exists, one can select the unlabeled sentence most similar to it in the same cluster. Then, pairs of concepts are collected from the selected sentences. The feature vectors of concept pairs will be used to retrain the relation classification model.

Unlike the *LDC*, the *UDP* method uses the label of each member to compute the purity of each cluster. Often, more than one pair of concepts exists in a sentence. When they have different relation types, the label of the sentence should be considered for the *UDP* method. One way would be to assign the most common relation type in the sentence.

In this dissertation, the instances were clustered based on cosine similarity. The sentence similarity can be calculated using various functions. Plank and Noord [200] selected new examples using Kullback-Leibler divergence, Jensen-Shannon divergence, Renyi divergence,

and other distance functions including cosine, Euclidean, and variational distance functions. They showed that using similarity functions to select new examples outperforms random data selection and even manual selection on dependency parsing accuracy.

6.2.4 Applying Other Active Learning Strategies

My instance selection methods compute the similarity between instances to get a diverse and balanced set of additional training instances. Applying other active learning approaches would extend the scope of this research.

The main goal of active learning is to collect beneficial examples from unlabeled data for humans to annotate for performance improvement. In Chapter 2, I discussed several popular active learning strategies that could be applied to select new training instances. The utility of each example can be assessed by various active learning strategies. In *uncertainty sampling* [146], the most uncertain example is selected. The instances can be selected based on the uncertainty (e.g., confidence score). In the case that multiple classification models are developed, the *query by committee* [231] strategy would be another good option. The instances the learners most disagree on can be considered as candidates for selection.

However, uncertainty sampling does not always result in the selection of informative instances. For example, an instance is often predicted less confidently when it contains features that are not in the training set. Adding this instance to the set of labeled data might not be beneficial when the unseen features are also not observed in the test set.

Another challenge is to determine the label of uncertain examples. Contrary to the assumption of an omniscient oracle in active learning, it can be more difficult in practice to determine the label of uncertain instances than other instances. This dissertation study has not focused on correcting label noise, but one direction for future research includes automatic label correction without the involvement of a human oracle. Clustering algorithms can be applied for label noise correction. For example, the label of an uncertain instance can be assigned based on the information of neighbor instances that are confidently labeled.

APPENDIX A

FREQUENT SECTION HEADERS IN EACH DATASET

This appendix lists the section headers that frequently appeared in the i2b2 Test data, and three specialty area data (cardiology, neurology, and orthopedics). Each table in this appendix shows the 40 most prevalent section headers in each dataset. The headers are sorted by the number of occurrence.

Table A.1: Section Headers Frequently Appearing in i2b2 Test

Section header	Count
HOSPITAL COURSE	180
HISTORY OF PRESENT ILLNESS	146
PHYSICAL EXAMINATION	136
PAST MEDICAL HISTORY	120
ALLERGIES	97
DISCHARGE MEDICATIONS	87
LABORATORY DATA	79
SOCIAL HISTORY	72
PRINCIPAL DIAGNOSIS	63
DISPOSITION	61
MEDICATIONS ON ADMISSION	60
MEDICATIONS ON DISCHARGE	54
DISCHARGE DIAGNOSES	48
CONDITION ON DISCHARGE	42
MEDICATIONS	38
PAST SURGICAL HISTORY	38
DISCHARGE DIAGNOSIS	38
IMPRESSION	36
ASSOCIATED DIAGNOSIS	36
DISCHARGE INSTRUCTIONS	35
FAMILY HISTORY	33
PRINCIPAL PROCEDURE	31
DISCHARGE CONDITION	27
SECONDARY DIAGNOSES	23
ACTIVITY	22
DISCHARGE DISPOSITION	22
OPERATIONS AND PROCEDURES	19
ABDOMEN	18
CHIEF COMPLAINT	18
FOLLOWUP	17
REVIEW OF SYSTEMS	17
TO DO / PLAN	16
ADDITIONAL COMMENTS	16
REASON FOR ADMISSION	16
HISTORY	16
HOSPITAL COURSE AND TREATMENT	15
BRIEF RESUME OF HOSPITAL COURSE	15
FOLLOW UP APPOINTMENT(S)	14
FINDINGS	14
ADMISSION DIAGNOSIS	14

Table A.2: Section Headers Frequently Appearing in Cardiology

Section header	Count
PHYSICAL EXAMINATION	74
ALLERGIES	69
PAST MEDICAL HISTORY	59
SOCIAL HISTORY	58
HISTORY OF PRESENT ILLNESS	56
FAMILY HISTORY	54
REVIEW OF SYSTEMS	53
MEDICATIONS	52
DISCHARGE MEDICATIONS	36
HOSPITAL COURSE	35
REASON FOR ADMISSION	34
DISCHARGE INSTRUCTIONS	32
IMPRESSION	31
LABORATORY DATA	25
EXTREMITIES	22
ABDOMEN	20
DISCHARGE DIAGNOSES	18
REASON FOR CONSULTATION	18
IMPRESSION AND PLAN	15
CHIEF COMPLAINT	14
PROCEDURE	14
PROCEDURES	13
ASSESSMENT AND PLAN	13
VITAL SIGNS	12
DISCHARGE DIAGNOSIS (ES)	11
CONCLUSION	11
DESCRIPTION OF OPERATION	11
PAST SURGICAL HISTORY	10
PAST MEDICAL/SURGICAL HISTORY	10
LABORATORY, RADIOGRAPHIC, AND OTHER DIAGNOSTIC STUDY FINDINGS	10
CURRENT MEDICATIONS	9
DESCRIPTION OF PROCEDURE	8
ACTIVITY	8
COMPLICATIONS	8
LABORATORY RESULTS	8
FOLLOWUP	8
HISTORY	8
CARDIOVASCULAR	8
DIAGNOSES	7
TRANSFER INSTRUCTIONS	7

Table A.3: Section Headers Frequently Appearing in Neurology

Section header	Count
HOSPITAL COURSE	82
REASON FOR ADMISSION	54
HISTORY OF PRESENT ILLNESS	47
DISCHARGE MEDICATIONS	46
DISCHARGE INSTRUCTIONS	45
PHYSICAL EXAMINATION	45
MEDICATIONS	41
SOCIAL HISTORY	39
ALLERGIES	38
FAMILY HISTORY	34
REVIEW OF SYSTEMS	33
FOLLOWUP	28
PAST MEDICAL HISTORY	27
IMPRESSION	25
TRANSFER INSTRUCTIONS	23
ACTIVITY	22
DISCHARGE DIAGNOSIS (ES)	21
DISPOSITION	21
CHIEF COMPLAINT	21
ASSESSMENT AND PLAN	19
TRANSFER MEDICATIONS	19
DISCHARGE DIAGNOSIS	19
SECONDARY DIAGNOSES	18
TRANSFER DIAGNOSIS (ES)	17
PAST MEDICAL/SURGICAL HISTORY	15
PROCEDURE	12
PROCEDURES	11
CONDITION	11
REASON FOR CONSULTATION	11
FOLLOW UP	11
PRINCIPAL DIAGNOSIS	11
DISCHARGE DIAGNOSES	9
ADMISSION DIAGNOSIS	9
ADMISSION DIAGNOSIS (ES)	9
BRIEF HISTORY	9
SECONDARY DIAGNOSIS (ES)	8
LABORATORY RESULTS	8
DIAGNOSIS	7
EXTREMITIES	7
PRINCIPAL PROCEDURES THIS ADMISSION	7

Table A.4: Section Headers Frequently Appearing in Orthopedics

Section header	Count
HOSPITAL COURSE	64
PROCEDURES	58
DISCHARGE INSTRUCTIONS	51
DESCRIPTION OF OPERATION	51
COMPLICATIONS	45
DISCHARGE MEDICATIONS	42
PROCEDURE	38
ANESTHESIA	34
POSTOPERATIVE DIAGNOSIS (ES)	34
DISCHARGE DIAGNOSIS (ES)	32
PREOPERATIVE DIAGNOSIS (ES)	31
INDICATIONS	30
DISCHARGE DIAGNOSIS	25
MEDICATIONS	22
ADMISSION DIAGNOSIS	21
DISPOSITION	20
HISTORY OF PRESENT ILLNESS	20
PREOPERATIVE DIAGNOSES	19
PREOPERATIVE DIAGNOSIS	19
REASON FOR ADMISSION	18
POSTOPERATIVE DIAGNOSES	18
POSTOPERATIVE DIAGNOSIS	18
PHYSICAL EXAMINATION	15
TITLE OF OPERATION	15
DISCHARGE DIAGNOSES	14
ALLERGIES	13
FINDINGS	13
PAST MEDICAL HISTORY	13
FOLLOWUP	13
ADMISSION HISTORY AND SUMMARY	12
INDICATIONS FOR SURGERY	11
ADMISSION HISTORY	10
SOCIAL HISTORY	10
PROCEDURES PERFORMED	10
CONDITION ON DISCHARGE	10
DESCRIPTION OF PROCEDURE	9
DISCHARGE CONDITION	9
ESTIMATED BLOOD LOSS	9
ADMISSION DIAGNOSIS (ES)	9
REVIEW OF SYSTEMS	9

APPENDIX B

SAMPLE SPECIALTY NOTES

Figure B.1 illustrates a cardiology note that is similar to the ones in our collection. Figure B.2 and Figure B.3 do the same for neurology and orthopedics, respectively. These example documents came from `http://www.mtsamples.com`.

CHIEF COMPLAINT: Palpitations.
CHEST PAIN / UNSPECIFIED ANGINA PECTORIS HISTORY: The patient relates the recent worsening of chronic chest discomfort. The quality of the pain is sharp and the problem started 2 years ago. Pain radiates to the back and condition is best described as severe. Patient denies syncope. Beyond baseline at present time. Past work up has included 24 hour Holter monitoring and echocardiography. Holter showed PVCs.
PALPITATIONS HISTORY: Palpitations - frequent, 2 x per week. No caffeine, no ETOH. + stress. No change with Inderal.
VALVULAR DISEASE HISTORY: Patient has documented mitral valve prolapse on echocardiography in 1992.
PAST MEDICAL HISTORY: No significant past medical problems. Mitral Valve Prolapse.
FAMILY MEDICAL HISTORY: CAD.
OB-GYN HISTORY: The patients last child birth was 1997. Para 3. Gravida 3.
SOCIAL HISTORY: Denies using caffeinated beverages, alcohol or the use of any tobacco products.
ALLERGIES: No known drug allergies/Intolerances.
CURRENT MEDICATIONS: Inderal 20 prn.
REVIEW OF SYSTEMS: Generally healthy. The patient is a good historian.
ROS Head and Eyes: Denies vision changes, light sensitivity, blurred vision, or double vision.
ROS Ear, Nose and Throat: The patient denies any ear, nose or throat symptoms.
ROS Respiratory: Patient denies any respiratory complaints, such as cough, shortness of breath, chest pain, wheezing, hemoptysis, etc.
ROS Gastrointestinal: Patient denies any gastrointestinal symptoms, such as anorexia, weight loss, dysphagia, nausea, vomiting, abdominal pain, abdominal distention, altered bowel movements, diarrhea, constipation, rectal bleeding, hematochezia.
ROS Genitourinary: Patient denies any genito-urinary complaints, such as hematuria, dysuria, frequency, urgency, hesitancy, nocturia, incontinence.
ROS Gynecological: Denies any gynecological complaints, such as vaginal bleeding, discharge, pain, etc.
ROS Musculoskeletal: The patient denies any past or present problems related to the musculoskeletal system.
ROS Extremities: The patient denies any extremities complaints.
ROS Cardiovascular: As per HPI. ...

Figure B.1: A Sample Cardiology Note

Description: Patient with a history of right upper pons and right cerebral peduncle infarction.
I had the pleasure of reevaluating Ms. A in our neurology clinic today for history of right upper pons and right cerebral peduncle infarction in April of 2008. Since her last visit in May of 2009, Ms. A stated that there has been no concern. She continues to complain of having mild weakness on the left leg at times and occasional off and on numbness in the left hand. She denied any weakness in the arm. She stated that she is ambulating with a cane. She denied any history of falls. Recently, she has also had repeat carotid Dopplers or further imaging studies of which we have no results of stating that she has further increased stenosis in her left internal carotid artery and there is a plan for surgery at Hospital with Dr. X. Of note, we have no notes to confirm that. Her daughter stated that she has planned for the surgery. Ms. A on today's office visit had no other complaints.
FAMILY HISTORY AND SOCIAL HISTORY: Reviewed and remained unchanged.
MEDICATIONS: List remained unchanged including Plavix, aspirin, levothyroxine, lisinopril, hydrochlorothiazide, Lasix, insulin and simvastatin.
ALLERGIES: She has no known drug allergies.
FALL RISK ASSESSMENT: Completed and there was no history of falls.
REVIEW OF SYSTEMS: Full review of systems again was pertinent for shortness of breath, lack of energy, diabetes, hypothyroidism, weakness, numbness and joint pain. Rest of them was negative.
PHYSICAL EXAMINATION:
Vital Signs: Today, blood pressure was 170/66, heart rate was 66, respiratory rate was 16, she weighed 254 pounds as stated, and temperature was 98.0.
General: She was a pleasant person in no acute distress.
HEENT: Normocephalic and atraumatic. No dry mouth. No palpable cervical lymph nodes. Her conjunctivae and sclerae were clear.
NEUROLOGICAL EXAMINATION: Remained unchanged.
Mental Status: Normal.
Cranial Nerves: Mild decrease in the left nasolabial fold.
Motor: There was mild increased tone in the left upper extremity. Deltoids showed 5-/5. The rest showed full strength. Hip flexion again was 5-/5 on the left. The rest showed full strength.
Reflexes: Reflexes were hypoactive and symmetrical.
Gait: She was mildly abnormal. No ataxia noted. Wide-based, ambulated with a cane. ...

Figure B.2: A Sample Neurology Note

PREOPERATIVE DIAGNOSIS: Achilles tendon rupture, left lower extremity.
POSTOPERATIVE DIAGNOSIS: Achilles tendon rupture, left lower extremity.
PROCEDURE PERFORMED: Primary repair left Achilles tendon.
ANESTHESIA: General.
COMPLICATIONS: None.
ESTIMATED BLOOD LOSS: Minimal.
TOTAL TOURNIQUET TIME: 40 minutes at 325 mmHg.
POSITION: Prone.
HISTORY OF PRESENT ILLNESS: The patient is a 26-year-old African-American male who states that he was stepping off a hilo at work when he felt a sudden pop in the posterior aspect of his left leg. The patient was placed in posterior splint and followed up at ABC orthopedics for further care.
PROCEDURE: After all potential complications, risks, as well as anticipated benefits of the above-named procedure were discussed at length with the patient, informed consent was obtained. The operative extremity was then confirmed with the patient, the operative surgeon, Department Of Anesthesia, and nursing staff. While in this hospital, the Department Of Anesthesia administered general anesthetic to the patient. The patient was then transferred to the operative table and placed in the prone position. All bony prominences were well padded at this time.
A nonsterile tourniquet was placed on the left upper thigh of the patient, but not inflated at this time. Left lower extremity was sterilely prepped and draped in the usual sterile fashion. Once this was done, the left lower extremity was elevated and exsanguinated using an Esmarch and the tourniquet was inflated to 325 mmHg and kept up for a total of 40 minutes. After all bony and soft tissue land marks were identified, a 6 cm longitudinal incision was made paramedial to the Achilles tendon from its insertion proximal. Careful dissection was then taken down to the level of the peritenon. Once this was reached, full thickness flaps were performed medially and laterally. Next, retractor was placed. All neurovascular structures were protected. A longitudinal incision was then made in the peritenon and opened up exposing the tendon. There was noted to be complete rupture of the tendon approximately 4 cm proximal to the insertion point. The plantar tendon was noted to be intact. The tendon was debrided at this time of hematoma as well as frayed tendon. Wound was copiously irrigated and dried. Most of the ankle appeared that there was sufficient tendon links in order to do a primary repair. Next #0 PDS on a taper needle was selected and a Krackow stitch was then performed. Two sutures were then used and tied individually _____ from the tendon. The wound was once again copiously irrigated and dried. ...

Figure B.3: A Sample Orthopedics Note

APPENDIX C

METAMAP SEMANTIC TYPES

This appendix lists MetaMap semantic types that were used for medical concept extraction. Refer to `http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt` for full semantic type list.

The first column and third column display mapping between abbreviations and the full semantic type names. The second column shows type unique identifier (TUI). The last column shows Prob(concept_type | semantic_type), the highest probability among mappings between MetaMap semantic category and three concept types (*problem*, *treatment*, and *test*).

Table C.1: MetaMap Semantic Types Used for Medical Concept Extraction

Abbr.	TUI	Full semantic type name	Prob.(%)
acab	T020	Acquired Abnormality	92.9
anab	T190	Anatomical Abnormality	89.8
antb	T195	Antibiotic	67.5
bact	T007	Bacterium	90.5
biof	T038	Biologic Function	60.0
bird	T012	Bird	81.8
carb	T118	Carbohydrate	67.6
celf	T043	Cell Function	60.0
cell	T025	Cell	95.7
cgab	T019	Congenital Abnormality	71.7
chvf	T120	Chemical Viewed Functionally	60.0
chvs	T104	Chemical Viewed Structurally	83.3
diap	T060	Diagnostic Procedure	88.0
dsyn	T047	Disease or Syndrome	88.5
enzy	T126	Enzyme	70.9
euka	T204	Eukaryote	54.3
horm	T125	Hormone	91.0
inpo	T037	Injury or Poisoning	77.2
lbpr	T059	Laboratory Procedure	89.6
lbtr	T034	Laboratory or Test Result	76.5
mbrt	T063	Molecular Biology Research Technique	75.0
medd	T074	Medical Device	64.6
mobd	T048	Mental or Behavioral Dysfunction	58.0
moft	T044	Molecular Function	60.7
neop	T191	Neoplastic Process	76.1
nnon	T114	Nucleic Acid, Nucleoside, or Nucleotide	68.5
nsba	T124	Neuroreactive Substance or Biogenic Amine	84.0
opco	T115	Organophosphorus Compound	100.0
orch	T109	Organic Chemical	70.9
orgm	T001	Organism	100.0
patf	T046	Pathologic Function	91.5
phsf	T039	Physiologic Function	57.6
phsu	T121	Pharmacologic Substance	72.7
sbst	T167	Substance	54.4
sosy	T184	Sign or Symptom	87.9
strd	T110	Steroid	65.4
tisu	T024	Tissue	64.3
topp	T061	Therapeutic or Preventive Procedure	65.4
vita	T127	Vitamin	77.6

APPENDIX D

PARTIAL MATCH RESULTS OF CONCEPT EXTRACTION

This appendix provides partial match results of medical concept extraction models. I used the evaluation script developed for the i2b2 Challenge to calculate recall, precision, and F_1 score. The semantic category is ignored, and a match is made if the reference standard text span has at least one word in common with the concept detected by the system. Table D.1 shows the performance of each MCE model based on recall, precision, and F_1 score. Table D.2 shows the performance of these ensembles.

Table D.1: Results of Individual MCE Models (Partial Match)

Model	Recall	Precision	F_1 score
i2b2 Test			
MetaMap	57.6	75.7	65.4
Rules (i2b2)	70.7	88.9	78.8
SVM (i2b2)	**94.4**	90.0	92.2
CRF-fwd (i2b2)	90.3	**95.5**	92.8
CRF-rev (i2b2)	90.8	95.3	**93.0**
CRF-rev ($i2b2_{180}$)	88.2	94.9	91.4
Cardiology			
MetaMap	57.1	73.5	64.2
Rules (i2b2)	71.1	81.5	76.0
Rules (Sp)	70.1	83.0	76.0
SVM (i2b2)	89.7	82.5	86.0
SVM (Sp)	**90.9**	82.5	86.5
CRF-fwd (i2b2)	84.9	88.4	86.6
CRF-rev (i2b2)	85.5	88.4	86.9
CRF-rev ($i2b2_{180}$)	83.4	88.3	85.8
CRF-fwd (Sp)	83.4	**90.5**	86.8
CRF-rev (Sp)	85.0	90.0	87.4
CRF-rev (i2b2+Sp)	87.5	89.6	**88.5**
Neurology			
MetaMap	59.0	69.6	63.9
Rules (i2b2)	69.0	83.4	75.5
Rules (Sp)	73.7	78.5	76.0
SVM (i2b2)	86.9	81.5	84.1
SVM (Sp)	**89.9**	80.4	84.9
CRF-fwd (i2b2)	82.5	88.5	85.4
CRF-rev (i2b2)	82.8	88.3	85.5
CRF-rev ($i2b2_{180}$)	81.1	88.4	84.6
CRF-fwd (Sp)	82.1	**89.5**	85.6
CRF-rev (Sp)	83.7	89.4	86.4
CRF-rev (i2b2+Sp)	86.2	89.1	**87.6**
Orthopedics			
MetaMap	58.7	68.4	63.2
Rules (i2b2)	63.6	78.0	70.1
Rules (Sp)	74.4	77.5	75.9
SVM (i2b2)	82.3	74.9	78.4
SVM (Sp)	**89.4**	77.6	83.1
CRF-fwd (i2b2)	72.7	86.3	78.9
CRF-rev (i2b2)	74.0	85.7	79.4
CRF-rev ($i2b2_{180}$)	71.2	85.4	77.7
CRF-fwd (Sp)	80.0	**89.9**	84.7
CRF-rev (Sp)	82.0	88.7	85.2
CRF-rev (i2b2+Sp)	84.0	88.5	**86.2**

Table D.2: Results of Ensemble Methods (Partial Match)

Model	**Cardiology**			**Neurology**			**Orthopedics**		
	Rec	**Pre**	**F**	**Rec**	**Pre**	**F**	**Rec**	**Pre**	**F**
CRF-rev (i2b2+Sp)	87.5	89.6	**88.5**	86.2	89.1	**87.6**	**84.0**	88.5	**86.2**
EasyAdapt	85.8	90.2	88.0	84.9	89.6	87.2	83.1	89.2	86.1
Voting (i2b2)	75.6	90.4	82.4	72.2	90.5	80.3	57.0	89.9	69.7
Voting (Sp)	69.6	**92.8**	79.6	65.9	**92.5**	77.0	60.2	**92.9**	73.1
Voting (i2b2+Sp)	**88.0**	88.5	88.2	**86.2**	88.1	87.1	79.9	88.3	83.9
Stacked (i2b2)	84.5	88.7	86.5	81.9	88.5	85.1	71.8	86.5	78.4
Stacked (Sp)	78.3	91.3	84.3	73.8	91.2	81.6	68.5	91.8	78.4
Stacked (i2b2+Sp)	80.0	91.1	85.2	77.8	91.5	84.1	70.7	91.6	79.8

APPENDIX E

SECTIONS FOR ASSERTION CLASSIFICATION

This appendix lists the section headers collected from the 2010 i2b2 challenge training set.

Table E.1: Section Headers Identified for Assertion Classification

Assertion type	**Section header**
Hypothetical	FOLLOWUP INSTRUCTIONS
	FOLLOW UP
	DISCHARGE ORDERS
	RECOMMENDATIONS
	MEDICATIONS AT THE TIME OF DISCHARGE
	POTENTIALLY SERIOUS INTERACTION
	TO DO / PLAN
	MEDICATIONS ON ADMISSION
	CONSULTANTS
	PRIMARY CARE PHYSICIAN
	MEDICATIONS AT TIME OF DISCHARGE
	MEDICATION
	DISPOSITION / PLAN ON DISCHARGE
Not patient	FAMILY HISTORY
	SOCIAL HISTORY
Conditional	Section headers containing ALLERGY or ALLERGIES

APPENDIX F

ASSERTION FEATURES CONTRIBUTION

This appendix displays results of each ablated classifier for each feature type. Table F.1 shows the detailed results of feature contribution grouped by assertion categories.

Table F.1: Features Contribution for Each Assertion Type (2nd Version)

Feature	Rec	Impact	Pre	Impact	F_1	Impact
Present						
All	98.0		95.1		96.5	
- Contextual	97.8	-0.2	94.8	-0.3	96.3	-0.2
- Lexical	97.5	-0.5	93.7	-1.4	95.6	-1.0
- Lexico-syntactic	98.0	0.0	94.9	-0.2	96.4	-0.1
- Syntactic	98.0	0.0	94.8	-0.2	96.4	-0.1
- Word vector	97.9	-0.1	94.9	-0.2	96.4	-0.1
Absent						
All	95.6		95.7		95.6	
- Contextual	94.9	-0.6	95.2	-0.4	95.1	-0.5
- Lexical	93.4	-2.2	93.8	-1.9	93.6	-2.0
- Lexico-syntactic	95.5	0.0	95.7	0.0	95.6	0.0
- Syntactic	95.2	-0.4	95.4	-0.2	95.3	-0.3
- Word vector	95.4	-0.2	95.4	-0.3	95.4	-0.2
Possible						
All	59.3		81.2		68.6	
- Contextual	59.3	0.0	79.0	-2.2	67.8	-0.8
- Lexical	52.2	-7.1	79.4	-1.9	63.0	-5.6
- Lexico-syntactic	56.7	-2.6	79.3	-2.0	66.1	-2.5
- Syntactic	57.5	-1.8	80.3	-1.0	67.0	-1.6
- Word vector	58.0	-1.4	80.9	-0.4	67.6	-1.0
Conditional						
All	26.9		78.0		40.0	
- Contextual	24.6	-2.3	70.0	-8.0	36.4	-3.6
- Lexical	6.4	-20.5	78.6	0.6	11.9	-28.1
- Lexico-syntactic	26.3	-0.6	79.0	1.0	39.5	-0.5
- Syntactic	26.3	-0.6	73.8	-4.2	38.8	-1.2
- Word vector	25.7	-1.2	78.6	0.6	38.8	-1.2
Hypothetical						
All	87.9		91.0		89.4	
- Contextual	86.9	-1.0	91.0	-0.1	88.9	-0.6
- Lexical	84.5	-3.4	88.5	-2.6	86.5	-3.0
- Lexico-syntactic	87.6	-0.3	91.4	0.4	89.5	0.0
- Syntactic	87.5	-0.4	91.7	0.6	89.5	0.1
- Word vector	87.6	-0.3	90.5	-0.6	89.0	-0.4
Not patient						
All	80.0		95.9		87.2	
- Contextual	62.8	-17.2	90.1	-5.8	74.0	-13.2
- Lexical	80.0	0.0	91.3	-4.5	85.3	-1.9
- Lexico-syntactic	80.7	0.7	95.9	0.0	87.6	0.4
- Syntactic	79.3	-0.7	95.8	0.0	86.8	-0.4
- Word vector	76.6	-3.5	96.5	0.6	85.4	-1.8

REFERENCES

[1] 2010 i2b2 / VA Challenge, *2010 i2b2 / VA challenge evaluation annotation file formatting*, 2010.

[2] ——, *2010 i2b2 / VA challenge evaluation assertion annotation guidelines*, 2010.

[3] ——, *2010 i2b2 / VA challenge evaluation concept annotation guidelines*, 2010.

[4] ——, *2010 i2b2 / VA challenge evaluation relation annotation guidelines*, 2010.

[5] A. Aamodt and E. Plaza, *Case-based reasoning: Foundational issues, methodological variations, and system approaches*, AI COMMUN., 7 (1994), pp. 39–59.

[6] S. Abney, *Bootstrapping*, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL-2002, Association for Computational Linguistics, 2002, pp. 360–367.

[7] American medical association, *AMA physician specialty group and codes*.

[8] G. Angeli, J. Tibshirani, J. Wu, and D. C. Manning, *Combining distant and partial supervision for relation extraction*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP-2014, Association for Computational Linguistics, 2014, pp. 1556–1567.

[9] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe, *Extraction of adverse drug effects from clinical records*, Stud. Health Technol. Inform., 160 (2010), pp. 739–43.

[10] A. R. Aronson and F.-M. Lang, *An overview of MetaMap: historical perspective and recent advances*, JAMIA, 17 (2010), pp. 229–236.

[11] M. Asahara and Y. Matsumoto, *Japanese named entity extraction with redundant morphological analysis*, in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1 of NAACL-2003, Association for Computational Linguistics, 2003, pp. 8–15.

[12] Automatic Content Extraction 2008, *English annotation guidelines for relations*, 2008.

[13] N. Bach and S. Badaskar, *A review of relation extraction*, tech. rep., Canergie Mellon University, School of Computer Science, 2007.

[14] M.-F. Balcan, A. Blum, and K. Yang, *Co-training and expansion: Towards bridging theory and practice*, in Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS-2004, MIT Press, 2004, pp. 89–96.

[15] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, *Open information extraction from the web*, in Proceedings of the 20th international joint conference on Artifical intelligence, vol. 7 of IJCAI-2007, Morgan Kaufmann Publishers Inc., 2007, pp. 2670–2676.

[16] C. A. Bejan, L. Vanderwende, F. Xia, and M. Yetisgen-Yildiz, *Assertion modeling and its role in clinical phenotype identification*, J. Biomed. Inform., 46 (2013), pp. 68–74.

[17] D. M. Bikel, R. Schwartz, and R. M. Weischedel, *An algorithm that learns what's in a name*, Mach. Learn., 34 (1999), pp. 211–231.

[18] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, *Automatic extraction of biological information from scientific text: Protein-protein interactions*, in Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, ISMB-1999, AAAI Press, 1999, pp. 60–67.

[19] J. Blitzer, M. Dredze, and F. Pereira, *Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification*, in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL-2007, Association for Computational Linguistics, June 2007, pp. 440–447.

[20] J. Blitzer, R. McDonald, and F. Pereira, *Domain adaptation with structural correspondence learning*, in Proceedings of the 2006 conference on empirical methods in natural language processing, EMNLP-2006, Association for Computational Linguistics, 2006, pp. 120–128.

[21] A. Blum and T. Mitchell, *Combining labeled and unlabeled data with co-training*, in Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT-1998, ACM, 1998, pp. 92–100.

[22] O. Bodenreider, *The unified medical language system (UMLS): integrating biomedical terminology*, Nucleic Acids Res., 32 (2004), pp. D267–D270.

[23] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, *Freebase: A collaboratively created graph database for structuring human knowledge*, in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD-2008, ACM, 2008, pp. 1247–1250.

[24] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, *Exploiting diverse knowledge sources via maximum entropy in named entity recognition*, in Proceedings of the Sixth Workshop on Very Large Corpora, 1998, pp. 152–160.

[25] R. Bossy, J. Jourde, A.-P. Manine, P. Veber, E. Alphonse, M. Van De Guchte, P. Bessières, and C. Nédellec, *BioNLP shared task - the bacteria track*, BMC. Bioinformatics, 13 (2012), p. S3.

[26] L. Breiman, *Stacked regressions*, Mach. Learn., 24 (1996), pp. 49–64.

[27] E. Brill and J. Wu, *Classifier combination for improved lexical disambiguation*, in Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1 of ACL-1998, Association for Computational Linguistics, 1998, pp. 191–195.

[28] S. Brody, R. Navigli, and M. Lapata, *Ensemble methods for unsupervised WSD*, in Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-2006, Association for Computational Linguistics, 2006, pp. 97–104.

[29] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.-P. Kriegel, *Extraction of semantic biomedical relations from text using conditional random fields*, BMC. Bioinformatics, 9 (2008), p. 207.

[30] R. C. Bunescu and R. J. Mooney, *A shortest path dependency kernel for relation extraction*, in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT-2005, Association for Computational Linguistics, 2005, pp. 724–731.

[31] B. L. Cairns, R. D. Nielsen, J. J. Masanz, J. H. Martin, M. S. Palmer, W. H. Ward, and G. K. Savova, *The MiPACQ clinical question answering system*, in Proceedings of the AMIA Annual Symposium, vol. 2011, 2011, pp. 171–180.

[32] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr., and T. M. Mitchell, *Coupled semi-supervised learning for information extraction*, in Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM-2010, ACM, 2010, pp. 101–110.

[33] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, *A simple algorithm for identifying negated findings and diseases in discharge summaries*, J. Biomed. Inform., 34 (2001), pp. 301–310.

[34] W. W. Chapman, D. Chu, and J. N. Dowling, *Context: An algorithm for identifying contextual features from clinical text*, in Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, 2007, pp. 81–88.

[35] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and Ö. Uzuner, *Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions*, JAMIA, 18 (2011), pp. 540–543.

[36] E. Charniak and M. Johnson, *Coarse-to-fine n-best parsing and maxent discriminative reranking*, in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, ACL-2005, Association for Computational Linguistics, 2005, pp. 173–180.

[37] C. Chelba and A. Acero, *Adaptation of maximum entropy capitalizer: Little data can help a lot*, Comput. Speech Lang., 20 (2006), pp. 382–399.

[38] J. Chen, D. Ji, C. L. Tan, and Z. Niu, *Relation extraction using label propagation based semi-supervised learning*, in Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-2006, Association for Computational Linguistics, 2006, pp. 129–136.

[39] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu, *Applying active learning to high-throughput phenotyping algorithms for electronic health records data*, JAMIA, 20 (2013), pp. e253–e259.

[40] L. T. Cheng, J. Zheng, G. K. Savova, and B. J. Erickson, *Discerning tumor status from unstructured mri reports–completeness of information in existing reports and utility of automated natural language processing*, J. Digit. Imaging, 23 (2010), pp. 119–132.

[41] N. A. Chinchor, *Overview of MUC-7/MET-2*, in Proceedings of the Seventh Message Understanding Conference (MUC-7), vol. 1, 1998.

[42] F. M. Chowdhury, A. B. Abacha, A. Lavelli, and P. Zweigenbaum, *Two different machine learning techniques for drug-drug interaction extraction*, in Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction, 2011, pp. 19–26.

[43] M. M. F. Chowdhury and A. Lavelli, *FBK-irst : A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information*, in Proceedings of the Seventh International Workshop on Semantic Evaluation, vol. 2 of SemEval-2013, Association for Computational Linguistics, 2013, pp. 351–355.

[44] L. M. Christensen, P. J. Haug, and M. Fiszman, *Mplus: A probabilistic medical language understanding system*, in Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3, BioMed-2002, Association for Computational Linguistics, 2002, pp. 29–36.

[45] H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, *Extraction of gene-disease relations from medline using domain dictionaries and machine learning.*, in Pacific Symposium on Biocomputing, vol. 11, 2006, pp. 4–15.

[46] C. Clark, J. Aberdeen, M. Coarr, D. Tresner-Kirsch, B. Wellner, A. Yeh, and L. Hirschman, *MITRE system for clinical assertion status classification*, JAMIA, 18 (2011), pp. 563–567.

[47] S. Clark, J. R. Curran, and M. Osborne, *Bootstrapping pos taggers using unlabelled data*, in Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003, CONLL-2003, Association for Computational Linguistics, 2003, pp. 49–55.

[48] A. M. Cohen and W. R. Hersh, *A survey of current work in biomedical text mining*, Brief. Bioinform., 6 (2005), pp. 57–71.

[49] D. Cohn, L. Atlas, and R. Ladner, *Improving generalization with active learning*, Mach. Learn., 15 (1994), pp. 201–221.

[50] M. Collins, *Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms*, in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10 of EMNLP-2002, Association for Computational Linguistics, 2002, pp. 1–8.

[51] ——, *Ranking algorithms for named-entity extraction: Boosting and the voted perceptron*, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL-2002, Association for Computational Linguistics, 2002, pp. 489–496.

[52] M. Collins and Y. Singer, *Unsupervised models for named entity classification*, in Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in

Natural Language Processing and Very Large Corpora, Association for Computational Linguistics, 1999, pp. 100–110.

[53] C. Cortes and V. Vapnik, *Support-vector networks*, Mach. Learn., 20 (1995), pp. 273–297.

[54] M. Craven and J. Kumlien, *Constructing biological knowledge bases by extracting information from text sources*, in Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press, 1999, pp. 77–86.

[55] A. Culotta and A. McCallum, *Reducing labeling effort for structured prediction tasks*, in Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2, AAAI-2005, AAAI Press, 2005, pp. 746–751.

[56] A. Culotta and J. Sorensen, *Dependency tree kernels for relation extraction*, in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL-2004, Association for Computational Linguistics, 2004, pp. 423–429.

[57] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, *GATE: A framework and graphical development environment for robust NLP tools and applications*, in Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, ACL-2002, 2002.

[58] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, *Getting more out of biomedical documents with gate's full lifecycle open source text analytics*, PLoS. Comput. Biol., 9 (2013), p. e1002854.

[59] J. R. Curran, T. Murphy, and B. Scholz, *Minimising semantic drift with mutual exclusion bootstrapping*, in Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, 2007, pp. 172–180.

[60] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka, *Tweet sentiment analysis with classifier ensembles*, Decis. Support Syst., 66 (2014), pp. 170–179.

[61] I. Dagan and S. P. Engelson, *Committee-based sampling for training probabilistic classifiers*, in Proceedings of the Twelfth International Conference on International Conference on Machine Learning, ICML-1995, Morgan Kaufmann Publishers Inc., 1995, pp. 150–157.

[62] H. Daumé, III, *Cross-task knowledge-constrained self training*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-2008, Association for Computational Linguistics, 2008, pp. 680–688.

[63] H. Daumé III, *Frustratingly easy domain adaptation*, in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL-2007, Association for Computational Linguistics, 2007, pp. 256–263.

[64] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu, *Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010*, JAMIA, 18 (2011), p. 557.

[65] J. De Las Rivas and C. Fontanillo, *Protein–protein interactions essentials: key concepts to building and analyzing interactome networks*, PLoS. Comput. Biol., 6 (2010), p. e1000807.

[66] F. Dell'Orletta, *Ensemble system for Part-of-Speech tagging*, in Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 2009.

[67] A. Demiriz, K. Bennett, and M. J. Embrechts, *Semi-supervised clustering using genetic algorithms*, in Proceedings of the Artificial Neural Networks in Engineering, ASME Press, 1999, pp. 809–814.

[68] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Stat. Soc., 39 (1977), pp. 1–38.

[69] T. G. Dietterich, *Ensemble methods in machine learning*, in Proceedings of the First International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.

[70] S. Doan, N. Collier, H. Xu, P. H. Duy, and T. M. Phuong, *Recognition of medication information from discharge summaries using ensembles of classifiers*, BMC. Med. Inform. Decis. Mak., 12 (2012), p. 36.

[71] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, *The automatic content extraction (ACE) program tasks, data, and evaluation*, in Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), European Language Resources Association (ELRA), May 2004.

[72] K. Doing-Harris, O. Patterson, S. Igo, and J. Hurdle, *Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts*, in Proceedings of the ACM International Workshop on Data and Text Mining in Biomedical Informatics, Oct-Nov 2013, pp. 9–12.

[73] P. Donmez and J. G. Carbonell, *Proactive learning: Cost-sensitive active learning with multiple imperfect oracles*, in Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM-2008, ACM, 2008, pp. 619–628.

[74] J. D'Souza and V. Ng, *Ensemble-based medical relation classification*, in Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, COLING-2014, Dublin City University and Association for Computational Linguistics, 2014, pp. 1682–1693.

[75] S. Džeroski and B. Ženko, *Is combining classifiers with stacking better than selecting the best one?*, Mach. Learn., 54 (2004), pp. 255–273.

[76] A. Ekbal and S. Saha, *Weighted vote-based classifier ensemble for named entity recognition: A genetic algorithm-based approach*, ACM. TALIP., 10 (2011), pp. 1–37.

[77] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, *Open information extraction from the web*, Commun. ACM., 51 (2008), pp. 68–74.

[78] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, *Unsupervised named-entity extraction from the web: An experimental study*, Artif. Intell., 165 (2005), pp. 91–134.

[79] A. Fader, S. Soderland, and O. Etzioni, *Identifying relations for open information extraction*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-2011, Association for Computational Linguistics, 2011, pp. 1535–1545.

[80] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, *LIBLINEAR: A library for large linear classification*, J. Mach. Learn. Res., 9 (2008), pp. 1871–1874.

[81] R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas, *The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text*, in Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task, Association for Computational Linguistics, 2010, pp. 1–12.

[82] D. Ferrucci and A. Lally, *Uima: an architectural approach to unstructured information processing in the corporate research environment*, Nat. Lang. Eng., 10 (2004), pp. 327–348.

[83] D. Ferrucci, A. Lally, K. Verspoor, and E. Nyberg, *Unstructured information management architecture (UIMA) version 1.0.* OASIS Standard, mar 2009.

[84] J. Finkel, S. Dingare, C. D. Manning, M. Nissim, B. Alex, and C. Grover, *Exploring the boundaries: gene and protein identification in biomedical text*, BMC. Bioinformatics, 6 (2005), p. S5.

[85] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, *Exploiting context for biomedical entity recognition: From syntax to the web*, in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics, 2004, pp. 88–91.

[86] R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, and S. Roukos, *A statistical model for multilingual entity detection and tracking*, in Proceedings of the Joint Conference of the North American Chapter of the Association for Computational Linguistics and Conference on Human Language Technologies, NAACL/HLT-2004, Association for Computational Linguistics, May 2 - May 7 2004, pp. 1–8.

[87] G. Foster and R. Kuhn, *Mixture-model adaptation for SMT*, in Proceedings of the second workshop on statistical machine translation, Association for Computational Linguistics, 2007, pp. 128–135.

[88] Y. Freund and R. E. Schapire, *Large margin classification using the perceptron algorithm*, Mach. Learn., 37 (1999), pp. 277–296.

[89] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, *A general natural-language text processor for clinical radiology*, JAMIA, 1 (1994), pp. 161–174.

[90] C. Friedman, P. Kra, and A. Rzhetsky, *Two biomedical sublanguages: a description based on the theories of Zellig Harris*, J. Biomed. Inform., 35 (2002), pp. 222–235.

[91] A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka, *Selective sampling for example-based word sense disambiguation*, Comput. Linguist., 24 (1998), pp. 573–597.

[92] K. Fundel, R. Küffner, and R. Zimmer, *RelEx—relation extraction using dependency parse trees*, Bioinformatics, 23 (2007), pp. 365–371.

[93] J. Gama and P. Brazdil, *Cascade generalization*, Mach. Learn., 41 (2000), pp. 315–343.

[94] A. B. Goldberg, X. Zhu, A. Singh, Z. Xu, and R. Nowak, *Multi-manifold semi-supervised learning*, in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 2009, pp. 169–176.

[95] Z. M. Grinspan, S. Banerjee, R. Kaushal, and L. Kern, *Physician specialty and variations in adoption of electronic health records*, Appl. Clin. Inform., 4 (2013), pp. 225–240.

[96] R. Grishman, *The NYU system for MUC-6, or where's the syntax?*, in Proceedings of the sixth message understanding conference (MUC-6), 1995.

[97] R. Grishman and B. Sundheim, *Message understanding conference-6: A brief history*, in Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING-1996, Association for Computational Linguistics, 1996, pp. 466–471.

[98] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, *Exploring various knowledge in relation extraction*, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL-2005, Association for Computational Linguistics, 2005, pp. 427–434.

[99] P. J. Haug, L. Christensen, M. Gundersen, B. Clemons, S. Koehler, and K. Bauer, *A natural language parsing system for encoding admitting diagnoses*, in Proceedings of the AMIA Annual Fall Symposium, 1997, pp. 814–818.

[100] P. J. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, and S. M. Huff, *Experience with a mixed semantic/syntactic parser*, in Proceedings of the Annual Symposium on Computer Application in Medical Care, 1995, pp. 284–288.

[101] L. He, Z. Yang, Z. Zhao, H. Lin, and Y. Li, *Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach*, PLoS. ONE, 8 (2013), p. e65814.

[102] M. A. Hearst, *Automatic acquisition of hyponyms from large text corpora*, in Proceedings of the 14th conference on Computational linguistics, Association for Computational Linguistics, 1992, pp. 539–545.

[103] D. T. Heinze, M. Morsch, R. Sheffer, M. Jimmink, M. Jennings, W. Morris, and A. Morsch, *Lifecode: A deployed application for automated medical coding*, AI. MAG., 22 (2001), p. 76.

[104] L. Hirschman, A. A. Morgan, and A. S. Yeh, *Rutabaga by any other name: extracting biological names*, J. Biomed. Inform., 35 (2002), pp. 247–259.

[105] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, *Overview of BioCreAtIvE: critical assessment of information extraction for biology*, BMC. Bioinformatics, 6 (2005), p. S1.

[106] S.-S. Ho and H. Wechsler, *Query by transduction*, IEEE. Trans. Pattern Anal. Mach. Intell., 30 (2008), pp. 1557–1571.

[107] J. R. Hobbs and E. Riloff, *Information extraction*, in Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC, 2010.

[108] A. Holub, P. Perona, and M. C. Burl, *Entropy-based active learning for object recognition*, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8.

[109] C.-J. Hsiao and E. Hing, *Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2013*, NCHS. Data Brief., no 143 (2014), pp. 1–8.

[110] C.-C. Huang and Z. Lu, *Community challenges in biomedical text mining over 10 years: success, failure and the future*, Brief. Bioinform., 17 (2016), pp. 132–144.

[111] H.-S. Huang, Y.-S. Lin, K.-T. Lin, C.-J. Kuo, Y.-M. Chang, B.-H. Yang, I.-F. Chung, and C.-N. Hsu, *High-recall gene mention recognition by unification of multiple backward parsing models*, in The Second BioCreative Challenge Evaluation Workshop (BioCreative II), 2007, pp. 109–111.

[112] R. Huang and E. Riloff, *Inducing domain-specific semantic class taggers from (almost) nothing*, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 275–285.

[113] H. Isozaki and H. Kazawa, *Efficient support vector classifiers for named entity recognition*, in Proceedings of the 19th International Conference on Computational Linguistics, vol. 1 of COLING-2002, Association for Computational Linguistics, 2002, pp. 1–7.

[114] J. Jiang and C. Zhai, *Instance weighting for domain adaptation in NLP*, in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL-2007, Association for Computational Linguistics, 2007, pp. 264–271.

[115] ——, *A systematic exploration of the feature space for relation extraction*, in Proceedings of the Joint Conference on Human Language Technologies and Conference of the North American Chapter of the Association for Computational Linguistics, HLT/NAACL-2004, Association for Computational Linguistics, 2007, pp. 113–120.

[116] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries*, JAMIA, 18 (2011), pp. 601–606.

[117] T. Joachims, *Transductive inference for text classification using support vector machines*, in Proceedings of the 16th International Conference on Machine Learning, ICML-1999, Morgan Kaufmann Publishers Inc., 1999, pp. 200–209.

[118] ——, *Transductive learning via spectral graph partitioning*, in Proceedings of the 20th International Conference on Machine Learning, ICML-2003, Amer Assn for Artificial, 2003, pp. 290–297.

[119] Joint Commission on Accreditation of Healthcare Organizations, *Hospital accreditation standards : 2006 HAS : accreditation policies, standards, elements of performance, scoring*, Oakbrook Terrace, IL. : Joint Commission Resources, 2006.

[120] R. Jones, *Learning to extract entities from labeled and unlabeled text*, PhD thesis, Carnegie Mellon University, 2005.

[121] N. Kambhatla, *Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations*, in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics, 2004, p. 22.

[122] N. Kang, Z. Afzal, B. Singh, E. M. van Mulligen, and J. A. Kors, *Using an ensemble system to improve concept extraction from clinical records*, J. Biomed. Inform., 45 (2012), pp. 423–428.

[123] G. Karypis, *CLUTO–a clustering toolkit*, tech. rep., University of Minnesota, Department of Computer Science, 2002.

[124] R. D. Kaufman, B. Sheehan, P. Stetson, R. A. Bhatt, I. A. Field, C. Patel, and M. J. Maisel, *Natural language processing–enabled and conventional data capture methods for input to electronic health records: A comparative usability study*, JMIR. Med. Inform., 4 (2016), p. e35.

[125] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, G. Zuccon, and J. Palotti, *Overview of the ShARe/CLEF eHealth evaluation lab 2014*, in Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, Springer International Publishing, 2014, pp. 172–191.

[126] H. Kilicoglu and S. Bergler, *Syntactic dependency based heuristics for biological event extraction*, in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, Association for Computational Linguistics, 2009, pp. 119–127.

[127] H. Kilicoglu, D. Demner-Fushman, T. C. Rindflesch, N. L. Wilczynski, and R. B. Haynes, *Towards automatic recognition of scientifically rigorous clinical research evidence*, JAMIA, 16 (2009), pp. 25–31.

[128] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, *Overview of BioNLP'09 shared task on event extraction*, in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, Association for Computational Linguistics, 2009, pp. 1–9.

[129] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, *Genia corpus–a semantically annotated corpus for bio-textmining*, Bioinformatics, 19 (2003), pp. i180–i182.

[130] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, *Introduction to the bio-entity recognition task at JNLPBA*, in Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, Association for Computational Linguistics, 2004, pp. 70–75.

[131] J.-D. Kim, Y. Wang, T. Takagi, and A. Yonezawa, *Overview of Genia event task in BioNLP shared task 2011*, in Proceedings of the BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics, 2011, pp. 7–15.

[132] S. Kim, J. Yoon, and J. Yang, *Kernel approaches for genic interaction extraction*, Bioinformatics, 24 (2008), pp. 118–126.

[133] Y. Kim, E. Riloff, and S. M. Meystre, *Improving classification of medical assertions in clinical notes*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 2 of ACL/HLT-2011, Association for Computational Linguistics, 2011, pp. 311–316.

[134] D. Klein, J. Smarr, H. Nguyen, and C. D. Manning, *Named entity recognition with character-level models*, in Proceedings of the seventh conference on Natural language learning, vol. 4 of HLT/NAACL-2003, Association for Computational Linguistics, 2003, pp. 180–183.

[135] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, *Overview of the protein-protein interaction annotation extraction task of BioCreative II*, Genome Biol., 9 (2008), p. S4.

[136] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, et al., *The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text*, BMC. Bioinformatics, 12 (2011), p. S3.

[137] S. Kripalani, F. LeFevre, C. Phillips, M. Williams, P. Basaviah, and D. Baker, *Deficits in communication and information transfer between hospital-based and primary care physicians: Implications for patient safety and continuity of care*, JAMA., 297 (2007), pp. 831–841.

[138] G. R. Krupka and K. Hausman, *IsoQuest: Description of the NetOwl (TM) extractor system as used in MUC-7*, in Proceedings of the seventh message understanding conference (MUC-7), vol. 7, 1998.

[139] T. Kudo and Y. Matsumoto, *Chunking with support vector machines*, in Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, NAACL-2001, Association for Computational Linguistics, 2001, pp. 1–8.

[140] L. I. Kuncheva and C. J. Whitaker, *Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy*, Mach. Learn., 51 (2003), pp. 181–207.

[141] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, in Proceedings of the Eighteenth International Conference on Machine Learning, ICML-2001, Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[142] T. Lavergne, O. Cappé, and F. Yvon, *Practical very large scale CRFs*, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL-2010, Association for Computational Linguistics, 2010, pp. 504–513.

[143] R. Leaman, R. Khare, and Z. Lu, *Challenges in clinical natural language processing for automated disorder normalization*, J. Biomed. Inform., 57 (2015), pp. 28–37.

[144] F. Leitner and M. Krallinger, *The FEBS Letters SDA corpus: A collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community*, FEBS. Lett., 584 (2010), pp. 4129–4130.

[145] D. D. Lewis and J. Catlett, *Heterogenous uncertainty sampling for supervised learning*, in Proceedings of the Eleventh International Conference on International Conference on Machine Learning, ICML-1994, Morgan Kaufmann Publishers Inc., 1994, pp. 148–156.

[146] D. D. Lewis and W. A. Gale, *A sequential algorithm for training text classifiers*, in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR-1994, Springer-Verlag New York, Inc., 1994, pp. 3–12.

[147] J. Li, Z. Zhang, X. Li, and H. Chen, *Kernel-based learning for biomedical relation extraction*, J. Assoc. Inf. Sci. Technol., 59 (2008), pp. 756–769.

[148] L. Li, W. Fan, D. Huang, Y. Dang, and J. Sun, *Boosting performance of gene mention tagging system by hybrid methods*, J. Biomed. Inform., 45 (2012), pp. 156–164.

[149] L. Li, R. Zhou, D. Huang, and W. Liao, *Integrating divergent models for gene mention tagging*, in Proceedings of the 2009 International Conference on Natural Language Processing and Knowledge Engineering, Sept 2009, pp. 1–7.

[150] K. P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng-treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, I. Kohane, et al., *Electronic medical records for discovery research in rheumatoid arthritis*, Arthritis Care Res., 62 (2010), pp. 1120–1127.

[151] C. Lin, T. Miller, D. Dligach, R. Plenge, E. Karlson, and G. Savova, *Maximal information coefficient for feature selection for clinical document classification*, in ICML Workshop on Machine Learning for Clinical Data, 2012.

[152] D. Lin, *Using collocation statistics in information extraction*, in Proceedings of the seventh message understanding conference (MUC-7), vol. 66, 1998.

[153] D. Lin and P. Pantel, *Dirt–discovery of inference rules from text*, in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001, pp. 323–328.

[154] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, *The unified medical language system*, Yearb. Med. Inform., (1993), pp. 41–51.

[155] N. Littlestone and M. Warmuth, *The weighted majority algorithm*, Inf. Comput., 108 (1994), pp. 212 – 261.

[156] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, *Understanding of internal clustering validation measures*, in Proceedings of 10th IEEE International Conference on Data Mining (ICDM), IEEE, 2010, pp. 911–916.

[157] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, *The Stanford CoreNLP natural language processing toolkit*, in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2014, pp. 55–60.

[158] E. Marsh and D. Perzanowski, *MUC-7 evaluation of ie technology: Overview of results*, in Proceedings of the seventh message understanding conference (MUC-7), vol. 20, 1998.

[159] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, *Open language learning for information extraction*, in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP/CoNLL-2012, Association for Computational Linguistics, 2012, pp. 523–534.

[160] A. McCallum, D. Freitag, and F. C. Pereira, *Maximum entropy markov models for information extraction and segmentation*, in Proceedings of the Seventeenth International Conference on Machine Learning, vol. 17, 2000, pp. 591–598.

[161] A. McCallum and W. Li, *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*, in Proceedings of the Seventh Conference on Natural Language Learning, CONLL-2003, Association for Computational Linguistics, 2003, pp. 188–191.

[162] A. McCallum and K. Nigam, *Employing EM and pool-based active learning for text classification*, in Proceedings of the Fifteenth International Conference on Machine Learning, ICML-1998, Morgan Kaufmann Publishers Inc., 1998, pp. 350–358.

[163] D. McClosky, E. Charniak, and M. Johnson, *Effective self-training for parsing*, in Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT/NAACL-2006, Association for Computational Linguistics, 2006, pp. 152–159.

[164] R. McDonald and F. Pereira, *Identifying gene and protein mentions in text using conditional random fields*, BMC. Bioinformatics, 6 (2005), p. S6.

[165] R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White, *Simple algorithms for complex relation extraction with applications to biomedical IE*, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL-2005, Association for Computational Linguistics, 2005, pp. 491–498.

[166] T. McIntosh and R. J. Curran, *Reducing semantic drift with bagging and distributional similarity*, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, 2009, pp. 396–404.

[167] S. M. Meystre, Y. Kim, J. Heavirland, J. Williams, B. E. Bray, and J. Garvin, *Heart failure medications detection and prescription status classification in clinical narrative documents*, Stud. Health Technol. Inform., 216 (2015), p. 609.

[168] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, *Extracting information from textual documents in the electronic health record: a review of recent research*, Yearb. Med. Inform., 35 (2008), p. 44.

[169] R. Mihalcea, *Co-training and self-training for word sense disambiguation.*, in CoNLL, 2004, pp. 33–40.

[170] A. Mikheev, M. Moens, and C. Grover, *Named entity recognition without gazetteers*, in Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, EACL-1999, Association for Computational Linguistics, 1999, pp. 1–8.

[171] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, CoRR., abs/1301.3781 (2013).

[172] S. Miller, H. Fox, L. Ramshaw, and R. Weischedel, *A novel use of statistical parsing to extract information from text*, in Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, NAACL-2000, Association for Computational Linguistics, 2000, pp. 226–233.

[173] T. A. Miller, D. Dligach, and G. K. Savova, *Active learning for coreference resolution*, in Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2012, pp. 73–81.

[174] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, *Distant supervision for relation extraction without labeled data*, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2 of ACL/IJCNLP-2009, Association for Computational Linguistics, 2009, pp. 1003–1011.

[175] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, *Protein–protein interaction extraction by leveraging multiple kernels and parsers*, Int. J. Med. Inform., 78 (2009), pp. e39–e46.

[176] Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii, *Evaluating contributions of natural language parsers to protein–protein interaction extraction*, Bioinformatics, 25 (2009), pp. 394–400.

[177] Y. Miyao and J. Tsujii, *Feature forest models for probabilistic HPSG parsing*, Comput. Linguist., 34 (2008), pp. 35–80.

[178] G. Moncecchi, J.-L. Minel, and D. Wonsever, *A survey of kernel methods for relation extraction*, in Workshop on NLP and Web-based Technologies, 2010.

[179] R. Morante, A. Liekens, and W. Daelemans, *Learning the scope of negation in biomedical texts*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-2008, Association for Computational Linguistics, 2008, pp. 715–724.

[180] P. G. Mutalik, A. Deshpande, and P. M. Nadkarni, *Use of general-purpose negation detection to augment concept indexing of medical documents*, JAMIA, 8 (2001), pp. 598–609.

[181] D. Nadeau and S. Sekine, *A survey of named entity recognition and classification*, LingInv., 30 (2007), pp. 3–26.

[182] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, *Natural language processing: an introduction*, JAMIA, 18 (2011), pp. 544–551.

[183] C. Nédellec, *Learning language in logic-genic interaction extraction challenge*, in Proceedings of the 4th Learning Language in Logic Workshop, LLL-2005, 2005, pp. 1–7.

[184] M. Netzer, G. Millonig, M. Osl, B. Pfeifer, S. Praun, J. Villinger, W. Vogel, and C. Baumgartner, *A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry*, Bioinformatics, 25 (2009), p. 941.

[185] M. Neves, *An analysis on the entity annotations in biological corpora*, F1000Res., 3 (2014), p. 96.

[186] C. L. Newman, D. J. Blake, and C. J. Merz, *UCI repository of machine learning databases*, 1998.

[187] V. Ng and C. Cardie, *Weakly supervised natural language learning without redundant views*, in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL-2003, Association for Computational Linguistics, 2003, pp. 94–101.

[188] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and S. Colquist, *Symbolic rule-based classification of lung cancer stages from free-text pathology reports*, JAMIA, 17 (2010), pp. 440–445.

[189] H. T. Nguyen and A. Smeulders, *Active learning using pre-clustering*, in Proceedings of the Twenty-first International Conference on Machine Learning, ICML-2004, ACM, 2004, pp. 623–630.

[190] T. T. V. Nguyen and A. Moschitti, *End-to-end relation extraction using distant supervision from external semantic repositories*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL/HLT-2011, Association for Computational Linguistics, 2011, pp. 277–282.

[191] R. D. Nielsen, J. Masanz, P. Ogren, W. Ward, J. H. Martin, G. Savova, and M. Palmer, *An architecture for complex clinical question answering*, in Proceedings of the 1st ACM International Health Informatics Symposium, ACM, 2010, pp. 395–399.

[192] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, *Text classification from labeled and unlabeled documents using EM*, Mach. Learn., 39 (2000), pp. 103–134.

[193] Z.-Y. Niu, D.-H. Ji, and C. L. Tan, *Word sense disambiguation using label propagation based semi-supervised learning*, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL-2005, Association for Computational Linguistics, 2005, pp. 395–402.

[194] J. Nivre and R. Mcdonald, *Integrating graphbased and transition-based dependency parsers*, in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL/HLT-2008, 2008, pp. 950–958.

[195] J. C. Park, H. S. Kim, and J.-J. Kim, *Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar*, in Pacific symposium on biocomputing, vol. 6, 2001, pp. 396–407.

[196] O. Patterson and J. F. Hurdle, *Document clustering of clinical narratives: a systematic study of clinical sublanguages*, in AMIA Annual Symposium Proceedings, vol. 2011, 2011, pp. 1099–1107.

[197] T. Pedersen, *A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation*, in Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL-2000, Association for Computational Linguistics, 2000, pp. 63–69.

[198] M. P. Perrone and L. N. Cooper, *When networks disagree: Ensemble methods for hybrid neural networks*, in Proceedings of Neural Networks for Speech and Image Processing, Chapman and Hall, 1993, pp. 126–142.

[199] D. Pierce and C. Cardie, *Limitations of co-training for natural language learning from large datasets*, in Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2001, pp. 1–9.

[200] B. Plank and G. van Noord, *Effective measures of domain similarity for parsing*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT-2011, Association for Computational Linguistics, 2011, pp. 1566–1576.

[201] J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran, *Robust relational parsing over biomedical literature: Extracting inhibit relations*, in Proceedings of the Pacific symposium on biocomputing, vol. 7, 2002, pp. 362–373.

[202] A. Qadir and E. Riloff, *Ensemble-based semantic lexicon induction for semantic tagging*, in Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM), SemEval-2012, Association for Computational Linguistics, 2012, pp. 199–208.

[203] L. Qian, G. Zhou, F. Kong, Q. Zhu, and P. Qian, *Exploiting constituent dependencies for tree kernel-based semantic relation extraction*, in Proceedings of the 22nd International Conference on Computational Linguistics, COLING-2008, Association for Computational Linguistics, 2008, pp. 697–704.

[204] L. R. Rabiner, *A tutorial on hidden markov models and selected applications in speech recognition*, in Proceedings of the IEEE, Ieee, 1989, pp. 257–286.

[205] N. F. Rajani, V. Viswanathan, Y. Bentor, and R. Mooney, *Stacked ensembles of information extractors for knowledge-base population*, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL/IJCNLP-2015, Association for Computational Linguistics, July 2015, pp. 177–187.

[206] D. Ravichandran and E. Hovy, *Learning surface text patterns for a question answering system*, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 41–47.

[207] K. Ravikumar, H. Liu, J. D. Cohn, M. E. Wall, and K. Verspoor, *Literature mining of protein-residue associations with graph rules learned through distant supervision*, J. Biomed. Semant., 3 (2012), pp. S2–S2.

[208] D. Rebholz-Schuhmann, A. J. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, K. Hornbostel, A. Kouznetsov, R. Witte, J. B. Laurila, C. J. Baker, C.-J. Kuo, S. Clematide, F. Rinaldi, R. Farkas, G. Móra, K. Hara, L. I. Furlong, M. Rautschka, M. L. Neves, A. Pascual-Montano, Q. Wei, N. Collier, M. F. M. Chowdhury, A. Lavelli, R. Berlanga, R. Morante, V. Van Asch, W. Daelemans, J. Marina, E. van Mulligen, J. Kors, and U. Hahn, *Assessment of NER solutions against the first and second CALBC Silver Standard Corpus*, J. Biomed. Semant., 2 (2011), p. S11.

[209] P. Resnik, *Selectional preference and sense disambiguation*, in Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Association for Computational Linguistics, 1997, pp. 52–57.

[210] P. S. Resnik, *Selection and Information: A Class-based Approach to Lexical Relationships*, PhD thesis, University of Pennsylvania, 1993.

[211] E. Riloff and J. Shepherd, *A corpus-based approach for building semantic lexicons*, in Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, 1997, pp. 117–124.

[212] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter, *EDGAR: Extraction of drugs, genes and relations from the biomedical literature*, in Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2000, pp. 517–528.

[213] B. Rink, S. Harabagiu, and K. Roberts, *Automatic extraction of relations between medical concepts in clinical texts*, JAMIA, 18 (2011), pp. 594–600.

[214] B. Roark and E. Charniak, *Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction*, in Proceedings of the 17th International Conference on Computational Linguistics, 1998, pp. 1110–1116.

[215] A. Roberts, R. Gaizauskas, and M. Hepple, *Extracting clinical relationships from patient narratives*, in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP '08, Association for Computational Linguistics, 2008, pp. 10–18.

[216] B. Rosario and M. A. Hearst, *Classifying semantic relations in bioscience texts*, in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL-2004, Association for Computational Linguistics, 2004, pp. 430–437.

[217] C. Rosenberg, M. Hebert, and H. Schneiderman, *Semi-supervised self-training of object detection models*, in Proceedings of the 7th IEEE Workshops on Application of Computer Vision, WACV/MOTION-2005, IEEE Computer Society, 2005, pp. 29–36.

[218] F. Rosenblatt, *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*, tech. rep., DTIC Document, 1961.

[219] N. Roy and A. McCallum, *Toward optimal active learning through sampling estimation of error reduction*, in Proceedings of the Eighteenth International Conference on Machine Learning, ICML-2001, Morgan Kaufmann Publishers Inc., 2001, pp. 441–448.

[220] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, *Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database*, Crit. Care Med., 39 (2011), pp. 952–960.

[221] E. F. T. K. Sang, *Noun phrase recognition by system combination*, in Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL-2000, Association for Computational Linguistics, 2000, pp. 50–55.

[222] A. Sarkar, *Applying co-training methods to statistical parsing*, in Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, NAACL-2001, Association for Computational Linguistics, 2001, pp. 1–8.

[223] V. Satopää, J. Albrecht, D. Irwin, and B. Raghavan, *Finding a "Kneedle" in a haystack: Detecting knee points in system behavior*, in Proceedings of the 31st International Conference on Distributed Computing Systems Workshops, IEEE, 2011, pp. 166–171.

[224] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, *Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications*, JAMIA, 17 (2010), pp. 507–513.

[225] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, *Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)*, in Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval 2013, Association for Computational Linguistics, 2013, pp. 341–350.

[226] I. Segura-Bedmar, P. Martinez, and D. Sanchez-Cisneros, *The 1st DDIExtraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts*, in Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011, Association for Computational Linguistics, 2011, pp. 1–9.

[227] B. Settles, *Biomedical named entity recognition using conditional random fields and rich feature sets*, in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics, 2004, pp. 104–107.

[228] ——, *Active learning literature survey*, tech. rep., Computer Sciences, University of Wisconsin-Madison, 2010.

[229] ——, *Active learning*, Morgan & Claypool Publishers, 2012.

[230] B. Settles and M. Craven, *An analysis of active learning strategies for sequence labeling tasks*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, Stroudsburg, PA, USA, 2008, Association for Computational Linguistics, pp. 1070–1079.

[231] H. S. Seung, M. Opper, and H. Sompolinsky, *Query by committee*, in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT-1992, ACM, 1992, pp. 287–294.

[232] M. S. Simpson and D. Demner-Fushman, *Biomedical text mining: A survey of recent progress*, in Mining text data, Springer, 2012, pp. 465–517.

[233] S. Sohn, J.-P. A. Kocher, C. G. Chute, and G. K. Savova, *Drug side effect extraction from clinical narratives of psychiatry and psychology patients*, JAMIA, 18 (2011), pp. i144–i149.

[234] S. Somashekhar, R. Kumarc, A. Rauthan, K. Arun, P. Patil, and Y. Ramya, *Double blinded validation study to assess performance of IBM artificial intelligence platform, Watson for oncology in comparison with Manipal multidisciplinary tumour board – First study of 638 breast cancer cases*, Cancer Res., 77 (2017), pp. S6–07–S6–07.

[235] Y. Song, E. Kim, G. G. Lee, and B.-k. Yi, *POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004*, in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics, 2004, pp. 100–103.

[236] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The general inquirer: A computer approach to content analysis*, MIT press, 1966.

[237] S. Strassel, A. Mitchell, and S. Huang, *Multilingual resources for entity extraction*, in Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15, Association for Computational Linguistics, 2003, pp. 49–56.

[238] A. Sun, R. Grishman, and S. Sekine, *Semi-supervised relation extraction with large-scale word clustering*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL/HLT-2011, Association for Computational Linguistics, 2011, pp. 521–529.

[239] L. L. Sun and X. Z. Wang, *A survey on active learning strategy*, in 2010 International Conference on Machine Learning and Cybernetics, vol. 1, 2010, pp. 161–166.

[240] S. Sun and D. R. Hardoon, *Active learning with extremely sparse labeled examples*, Neurocomput., 73 (2010), pp. 2980–2988.

[241] W. Sun, A. Rumshisky, and Ö. Uzuner, *Evaluating temporal relations in clinical text: 2012 i2b2 challenge*, JAMIA, 20 (2013), pp. 806–813.

[242] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, and G. Zuccon, *Overview of the ShARe/CLEF eHealth evaluation lab 2013*, in Proceedings of the 4th International Conference of the CLEF Initiative, CLEF 2013, 2013, pp. 212–231.

[243] M. Surdeanu, *Overview of the TAC2013 knowledge base population evaluation: English slot filling and temporal slot filling*, in Proceedings of the 6th Text Analysis Conference (TAC2013), 2013.

[244] M. Surdeanu and H. Ji, *Overview of the english slot filling track at the TAC2014 knowledge base population evaluation*, in Proceedings of the 7th Text Analysis Conference (TAC2014), 2014.

[245] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, *Multi-instance multi-label learning for relation extraction*, in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP/CoNLL-2012, Association for Computational Linguistics, 2012, pp. 455–465.

[246] S. Takamatsu, I. Sato, and H. Nakagawa, *Reducing wrong labels in distant supervision for relation extraction*, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL-2012, Association for Computational Linguistics, 2012, pp. 721–729.

[247] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, *Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features*, BMC. Med. Inform. Decis. Mak., 13 (2013), p. S1.

[248] M. Thelen and E. Riloff, *A bootstrapping method for learning semantic lexicons using extraction pattern contexts*, in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, EMNLP-2002, Association for Computational Linguistics, 2002, pp. 214–221.

[249] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, *Automatic extraction of protein interactions from scientific abstracts*, in Pacific symposium on biocomputing, vol. 5, 2000, pp. 538–549.

[250] P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser, *Relation extraction for drug-drug interactions using ensemble learning*, in Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction, 2011, pp. 11–18.

[251] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser, *A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature*, PLoS. Comput. Biol., 6 (2010), pp. 1–19.

[252] K. M. Ting and I. H. Witten, *Issues in stacked generalization*, J. Artif. Int. Res., 10 (1999), pp. 271–289.

[253] E. F. Tjong Kim Sang, *Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition*, in Proceedings of the 6th Conference on Natural Language Learning, COLING-2002, Association for Computational Linguistics, 2002, pp. 1–4.

[254] E. F. Tjong Kim Sang and F. De Meulder, *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition*, in Proceedings of the Seventh Conference on Natural Language Learning at HLT/NAACL-2003, CONLL-2003, Association for Computational Linguistics, 2003, pp. 142–147.

[255] S. Tong and D. Koller, *Support vector machine active learning with applications to text classification*, J. Mach. Learn. Res., 2 (2001), pp. 45–66.

[256] W. Tong and R. Jin, *Semi-supervised learning by mixed label propagation*, in Proceedings of the 22nd National Conference on Artificial Intelligence, AAAI-2007, AAAI Press, 2007, pp. 651–656.

[257] O. Tuason, L. Chen, H. Liu, J. Blake, and C. Friedman, *Biological nomenclatures: A source of lexical knowledge and ambiguity*, in Biocomputing 2004, World Scientific, 2003, pp. 238–249.

[258] K. Tumer and J. Ghosh, *Error correlation and error reduction in ensemble classifiers*, Conn. Sci., 8 (1996), pp. 385–404.

[259] Ö. Uzuner, *Recognizing obesity and comorbidities in sparse data*, JAMIA, 16 (2009), pp. 561–570.

[260] Ö. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South, *Evaluating the state of the art in coreference resolution for electronic medical records*, JAMIA, 19 (2012), pp. 786–791.

[261] Ö. Uzuner, I. Goldstein, Y. Luo, and I. Kohane, *Identifying patient smoking status from medical discharge records*, JAMIA, 15 (2008), pp. 14–24.

[262] Ö. Uzuner, I. Solti, and E. Cadag, *Extracting medication information from clinical text*, JAMIA, 17 (2010), pp. 514–518.

[263] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text*, JAMIA, 18 (2011), pp. 552–556.

[264] Ö. Uzuner, X. Zhang, and T. Sibanda, *Machine learning and rule-based approaches to assertion classification*, JAMIA, 16 (2009), pp. 109–115.

[265] H. van Halteren, J. Zavrel, and W. Daelemans, *Improving data driven wordclass tagging by system combination*, in Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL-1998, Association for Computational Linguistics, 1998, pp. 491–497.

[266] C. J. Van Rijsbergen, *Information retrieval*, Butterworth-Heinemann, 2nd edition ed., 1979.

[267] V. N. Vapnik, *Statistical learning theory*, vol. 3, John Wiley, New York, 1998.

[268] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, *The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes*, BMC. Bioinformatics, 9 (2008), p. S9.

[269] X. Wan, *Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-2008, Association for Computational Linguistics, 2008, pp. 553–561.

[270] ——, *Co-training for cross-lingual sentiment classification*, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL-2009, Association for Computational Linguistics, 2009, pp. 235–243.

[271] C. Wang and J. Fan, *Medical relation extraction with manifold models*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL-2014, Association for Computational Linguistics, 2014, pp. 828–838.

[272] M. Wang and X.-S. Hua, *Active learning in multimedia annotation and retrieval: A survey*, ACM. TIST., 2 (2011), p. 10.

[273] S.-Q. Wang, J. Yang, and K.-C. Chou, *Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition*, J. Theor. Biol., 242 (2006), pp. 941–946.

[274] W. Wang and Z.-H. Zhou, *On multi-view active learning and the combination with semi-supervised learning*, in Proceedings of the 25th International Conference on Machine Learning, ICML-2008, ACM, 2008, pp. 1152–1159.

[275] D. Widdows and B. Dorow, *A graph model for unsupervised lexical acquisition*, in Proceedings of the 19th International Conference on Computational Linguistics, COLING-2002, Association for Computational Linguistics, 2002, pp. 1–7.

[276] D. H. Wolpert, *Stacked generalization*, Neural Netw., 5 (1992), pp. 241–259.

[277] F. Wu and D. S. Weld, *Open information extraction using Wikipedia*, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL-2010, Association for Computational Linguistics, 2010, pp. 118–127.

[278] M. Xiao and Y. Guo, *Domain adaptation for sequence labeling tasks with a probabilistic language adaptation model*, in Proceedings of the 30th Int. Conf. on Machine Learning, 2013, pp. 293–301.

[279] F. Xu, H. Uszkoreit, and H. Li, *A seed-driven bottom-up machine learning framework for extracting relations of various complexity*, in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL-2007, Association for Computational Linguistics, 2007, pp. 584–591.

[280] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, *Event extraction from biomedical papers using a full parser.*, in Pacific Symposium on Biocomputing, vol. 6, 2001, pp. 408–419.

[281] L. Yao, S. Riedel, and A. McCallum, *Collective cross-document relation extraction without labelled data*, in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP-2010, Association for Computational Linguistics, 2010, pp. 1013–1023.

[282] D. Yarowsky, *Unsupervised word sense disambiguation rivaling supervised methods*, in Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, ACL-1995, Association for Computational Linguistics, 1995, pp. 189–196.

[283] A. Yeh, *More accurate tests for the statistical significance of result differences*, in Proceedings of the 18th Conference on Computational Linguistics, COLING-2000, Association for Computational Linguistics, 2000, pp. 947–953.

[284] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman, *BioCreAtIvE Task 1A: gene mention finding evaluation*, BMC. Bioinformatics, 6 (2005), p. S2.

[285] D. Zelenko, C. Aone, and A. Richardella, *Kernel methods for relation extraction*, J. Mach. Learn. Res., 3 (2003), pp. 1083–1106.

[286] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system*, BMC. Med. Inform. Decis. Mak., 6 (2006), p. 30.

[287] M. Zhang, J. Zhang, and J. Su, *Exploring syntactic features for relation extraction using a convolution tree kernel*, in Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT/NAACL-2006, Association for Computational Linguistics, 2006, pp. 288–295.

[288] M. Zhang, J. Zhang, J. Su, and G. Zhou, *A composite kernel to extract relations between entities with both flat and structured features*, in Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-2006, Association for Computational Linguistics, 2006, pp. 825–832.

[289] T. Zhang and F. J. Oles, *A probability analysis on the value of unlabeled data for classification problems*, in Proceedings of the 17th International Conference on International Conference on Machine Learning, ICML-2000, Morgan Kaufmann, San Francisco, CA, 2000, pp. 1191–1198.

[290] Z. Zhang, *Weakly-supervised relation classification for information extraction*, in Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM-2004, ACM, 2004, pp. 581–588.

[291] S. Zhao, *Named entity recognition in biomedical texts using an HMM model*, in Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, Association for Computational Linguistics, 2004, pp. 84–87.

[292] S. Zhao and R. Grishman, *Extracting relations with integrated information using kernel methods*, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL-2005, Association for Computational Linguistics, 2005, pp. 419–426.

[293] J. Zheng, W. W. Chapman, T. A. Miller, C. Lin, R. S. Crowley, and G. K. Savova, *A system for coreference resolution for the clinical narrative*, JAMIA, 19 (2012), pp. 660–667.

[294] D. Zhou and Y. He, *Extracting interactions between proteins from the literature*, J. Biomed. Inform., 41 (2008), pp. 393–407.

[295] G. Zhou, D. Shen, J. Zhang, J. Su, and S. Tan, *Recognition of protein/gene names from text using an ensemble of classifiers*, BMC. Bioinformatics, 6 (2005), pp. S7–S7.

[296] G. Zhou and J. Su, *Named entity recognition using an HMM-based chunk tagger*, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL-2002, Association for Computational Linguistics, 2002, pp. 473–480.

[297] G. Zhou and J. Su, *Exploring deep knowledge resources in biomedical name recognition*, in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics, 2004, pp. 96–99.

[298] G. Zhou, M. Zhang, D. Ji, and Q. Zhu, *Tree kernel-based relation extraction with context-sensitive structured parse tree information*, in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP/CoNLL-2007, Association for Computational Linguistics, 2007, pp. 728–736.

[299] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, *Approaches to text mining for clinical medical records*, in Proceedings of the 2006 ACM Symposium on Applied Computing, SAC-2006, ACM, 2006, pp. 235–239.

[300] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, *Multi-instance learning by treating instances as non-I.I.D. samples*, in Proceedings of the 26th Annual International Conference on Machine Learning, ICML-2009, ACM, 2009, pp. 1249–1256.

[301] X. Zhu, *Semi-supervised learning literature survey*, Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[302] X. Zhu, C. Cherry, S. Kiritchenko, J. Martin, and B. de Bruijn, *Detecting concept relations in clinical text: Insights from a state-of-the-art model*, J. Biomed. Inform., 46 (2013), pp. 275–285.

[303] X. Zhu and Z. Ghahramani, *Learning from labeled and unlabeled data with label propagation*, tech. rep., Carnegie Mellon University, 2002.

[304] X. Zhu and A. B. Goldberg, *Introduction to semi-supervised learning*, Morgan & Claypool Publishers, 2009.

[305] X. Zhu, J. Lafferty, and Z. Ghahramani, *Combining active learning and semi-supervised learning using gaussian fields and harmonic functions*, in ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, 2003.

[306] Q. Zou, W. W. Chu, C. A. Morioka, G. H. Leazer, and H. Kangarloo, *IndexFinder: a method of extracting key concepts from clinical texts for indexing*, in AMIA Annual Symposium Proceedings, 2003, pp. 763–767.