**Methods and Techniques for Clinical Text Modeling and Analytics**

A Thesis

Submitted to the Faculty

of

Drexel University

by

Yuan Ling

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

February 2017

# Dedications

This thesis is dedicated to my family

whose love and support made this work possible

# Acknowledgments

I have received many support and encouragement from many people in the past five years.

First and foremost, I want to thank my supervisor Dr. Yuan An, whose generous guidance and support made it possible for me to complete my thesis. Dr. An devoted his precious time and effort to help me improve paper writing, research presentation, and critical thinking. His insight and vision in healthcare lead me to pursue a career in the healthcare industry. I am grateful to be his student.

I would like to thank my co-supervisor Dr. Tony Xiaohua Hu. Dr. Hu provided a flexible environment for me to explore different areas of research, and always support me with valuable resources and professional suggestions. He inspired me to work harder in pursuing a Ph.D. degree. I am grateful to be his student.

I would like to thank my dissertation committee of Dr. Weimao Ke, Dr. Ali Shokoufandeh, Dr. Jeffrey Headd for their support and insight throughout my research. They gave me many suggestions on my research work. I also want to thank Dr. Sadid Hasan and Dr. Oladimeji Farri from Philips Research North America for their help in my research.

I would like to thank my fellow graduate students and friends: Mengwen Liu, Wanying Ding, Yue Shang, Shi Ye, Xuelian Pan, Yetian Fan and other friends for all the help and support they provided.

Finally, I would like to thank my family, for their unconditional love.

# Table of Contents

# List of Tables

# List of Figures

# Abstract

Methods and Techniques for Clinical Text Modeling and Analytics
Yuan Ling
-

Nowadays, a large portion of clinical data only exists in free text. The wide adoption of Electronic Health Records (EHRs) has enabled the increases in accessing to clinical documents, which provide challenges and opportunities for clinical Natural Language Processing (NLP) researchers. Given free-text clinical notes as input, an ideal system for clinical text understanding should have the ability to support clinical decisions. At corpus level, the system should recommend similar notes based on disease or patient types, and provide medication recommendation, or any other type of recommendations, based on patients' symptoms and other similar medical cases. At document level, it should return a list of important clinical concepts. Moreover, the system should be able to make diagnostic inferences over clinical concepts and output diagnosis. Unfortunately, current work has not systematically studied this system.

This study focuses on developing and applying methods/techniques in different aspects of the system for clinical text understanding, at both corpus and document level. We deal with two major research questions:

First, we explore the question of *How to model the underlying relationships from clinical notes at corpus level?*

Documents clustering methods can group clinical notes into meaningful clusters, which can assist physicians and patients to understand medical conditions and diseases from clinical notes. We use Nonnegative Matrix Factorization (NMF) and Multi-view NMF to cluster clinical notes based on extracted medical concepts. The clustering results display latent patterns existed among clinical notes. Our method provides a feasible way to visualize a corpus of clinical documents. Based on extracted concepts, we further build a symptom-medication (Symp-Med) graph to model the Symp-Med relations in clinical notes corpus. We develop two Symp-Med matching algorithms to predict

and recommend medications for patients based on their symptoms.

Second, we want to solve the question of *How to integrate structured knowledge with unstructured text to improve results for Clinical NLP tasks?*

On the one hand, the unstructured clinical text contains lots of information about medical conditions. On the other hand, structured Knowledge Bases (KBs) are frequently used for supporting clinical NLP tasks. We propose graph-regularized word embedding models to integrate knowledge from both KBs and free text. We evaluate our models on standard datasets and biomedical NLP tasks, and results showed encouraging improvements on both datasets. We further apply the graph-regularized word embedding models and present a novel approach to automatically infer the most probable diagnosis from a given clinical narrative.

# Chapter 1: Introduction

## 1.1 Background

A great deal of effort has been put into improving health care in different aspects [4]. The adoption of Electronic Health Records (EHRs) is one of the ways to improve Healthcare. For example, clinical NLP tools [5, 6] are built based on EHR data to automatically trigger alerts and reminders for situations that require actions from physicians. EHRs[1] are the electronic version of patients' medical history, that are maintained by healthcare providers over time. Nowadays, EHRs are widely adopted by hospitals in the United States. Statistics from Figure 1.1 display the increasing percentage of non-federal acute care hospitals with the adoption of EHR systems over the years 2008 - 2015[2]. In addition to the increasing EHR adoption rate, the trends also show that there is an increasing use of advanced functionality for EHR systems. More and more hospitals are using EHRs with Clinical Notes and comprehensive EHRs with extra advanced functionality, such as decision support based on clinical guidelines, drug-drug interactions, drug allergy results, and etc[3].

EHRs make clinical notes digitalized and facilitate the way of sharing unstructured clinical notes with patients, which brings lots of benefits for both patients and physicians[4]. With access to clinical notes, patients will be able to take ownership of their own health, get more communication with healthcare providers, and understand their medical conditions better. For physicians and hospitals, digitalized clinical notes can be used as tools for them to find evidence for their decision-making process. Clinical notes can also be utilized by researchers to conduct research on clinical decision support. Researchers usually have limited access to EHR data due to the patient privacy protection.

As the development of de-identification techniques for EHRs [7, 8] and the guidance issued for

[1] https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html?redirect=/EhealthRecords/

[2] https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php#figure5

[3] https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php#appendix

[4] http://health.usnews.com/health-news/best-hospitals/articles/2015/10/15/hospitals-are-moving-slowly-to-electronic-medical-records

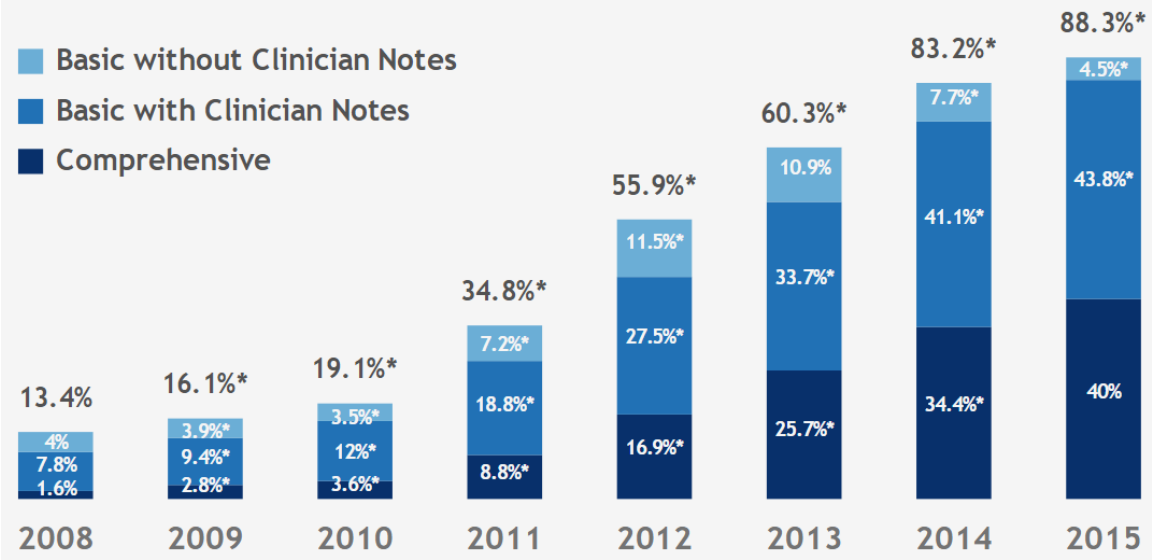



Figure 1.1 Percent of non-federal acute care hospitals with adoption of EHR systems by level of functionality: 2008 - 2015 (Statistics from Henry et al [1]).

the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[5], de-identified EHR data become more available to researchers. The lack of reproducibility problem [9] existed in related research will be allevaited since researchers can conduct experiments on same datasets.

There are two common ways for researchers to get unstructured clinical notes from EHR system. The first one is to obtain data through collaborations with hospitals [4]; and the second one is to get datasets through clinical NLP shared tasks for research purpose. For example, the i2b2 project (informatics for integration of biology and the bedside)[6] is an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. It creates shared tasks enable researchers to use existing clinical data for discovery research. Shared tasks provide annotated datasets and common evaluation metrics for participants [10].

## 1.2 Motivation

The increasing access to unstructured clinical notes brings chanllenges and opportunities [11] for research in Natural Language Processing (NLP) [12] and Information Retrieval (IR) areas to provide advanced techniques and tools for better understanding of clinical text. NLP and IR techniques

[5] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/
[6] https://www.i2b2.org/

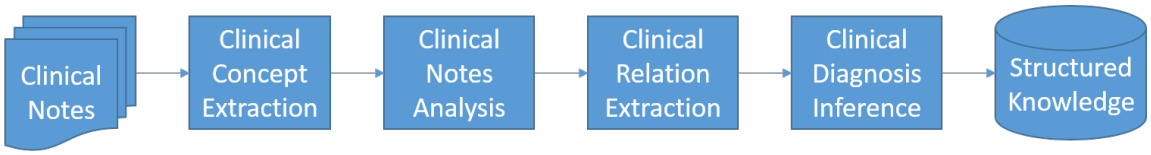



Figure 1.2 A Work Flow for Clinical Text Understanding

are applied to clinical text understanding for different type of tasks [13], such as clinical notes de-identification [14], clinical concept extraction [15, 16, 17], clinical relation extraction [18], biomedical literature retrieval [19], clinical question answering [20], and etc.

Such NLP and IR based systems build the foundation for clinical text analysis, which would satisfy the needs coming from both physicians and patients. Patients usually want to figure out their medical conditions from clinical notes, some general questions they want to get answers from clinical notes would be *what are my symptoms? how to treat the symptoms? what's the diagnosis for me? What are the conditions for other people with similar symptoms as me?* and all other related questions. Physicians can use patients' clinical notes for other purposes. For example, finding related medical cases or biomedical literature as evidence to support their decision-making process. They will ask questions such as *what's the drug choice for these symptoms? what're the causes for this symptom? How common is the disease?* and etc.

In this thesis, we study the problem of modeling and analyzing clinical notes. We develop novel methods and techniques for better understanding clinical text, in order to answer partial of these questions raised by physicians or patients. As displayed in Figure 1.2, different modules are included: clinical concept extraction, clinical notes analysis, clinical relation extraction, and clinical diagnosis inference.

Concept extraction from clinical notes is the foundation for clinical text understanding. Over the last decades, different methods and tools are developed for biomedical concept extraction [21, 16, 22, 23]. Extracting clinical concepts requires different types of systems designed for different types of clinical text. Clinical concepts, such as *finding*, *treatment*, *test*, *disease*, *genetic names*, and etc., from clinical text can be used for answering the aforementioned questions *what are my symptoms?*

*what's the diagnosis for me?*

Clinical notes clustering at corpus level provides solutions for questions like *What are the conditions for other people with similar symptoms as me? How common is the disease?* Clinical notes clustering requires different types of features from documents, such as word features, clinical concepts, risk factors of diseases, and etc. We explore to build clinical concepts enhanced document clustering methods for clinical notes clustering.

Clinical relation extraction refers as the classification of relationships between clinical concepts. For example, modeling symptom and medication relationships from clinical corpus [24] can help answering the questions of *how to treat the symptoms?* and *what's the drug choice for these symptoms?*

Clinical diagnosis inference is the problem of automatically inferring the most probable diagnosis from a given clinical narrative. Clinical diagnosis inference is the research work to answer the question as *what's the diagnosis?*

The motivation to answer aforementioned questions makes it desirable to develop NLP/IR methods and build tools for clinical text understanding.

## 1.3 Research Questions

Motivated by the general questions raised by patients and physicians' needs, we systematically study methods and techniques to achieve better clinical text understanding.

At corpus level, we want to explore the questions of *What kinds of relationships we can infer from a corpus of clinical notes? and How to model the relationships at corpus level?*

First, we want to explore the intrinsic relationships among clinical notes. Compared with general document clustering method, we incorporate extracted clinical concepts for clinical document clustering. We also want to use concept enhanced clinical document clustering to analyze and visualize the risk factors for heart disease in the diabetic population. We need to integrate multiple risk factors with various attributes into uniform feature representations, and clusters patients' data from multiple aspects. Second, we want to explore the symptoms/medications relationships exist in clinical notes corpus. Taking symptoms as input, we want to predict and recommend medications

for symptoms.

At document level, we want to answer the question of *How to integrate structured knowledge with unstructured text to improve results for Clinical NLP tasks?*

Word embedding in the NLP area has attracted increasing attention in recent years. The continuous bag-of-words model (CBOW) and the continuous Skip-gram model (Skip-gram) have been developed to learn distributed representations of words from a large amount of unlabeled text data. Besides, Knowledge Bases (KBs) are useful resources for supporting clinical NLP tasks. We explore the idea of integrating KBs with unstructured text and addressing the limitations of word embedding models when applied to clinical NLP tasks. There is a growing number of studies on applying word embedding models to biomedical NLP tasks. Overall, they focus on evaluating word embedding features and parameters trained on the biomedical corpus. There is little work on integrating KBs with word embedding models for biomedical NLP tasks.

## 1.4 Contributions

Given free-text notes as input, an ideal system for clinical text understanding should have the ability to support clinical decisions. At corpus level, the system should recommend similar notes based on disease or patient types, and provide medication recommendation, or any other types of recommendation, based on patients' symptoms and other similar medical cases. At document level, it should return a list of important clinical concepts. Moreover, the system should be able to make diagnostic inferences over clinical concepts and output diagnosis. Unfortunately, current work has not systematically studied this system. In our thesis, we develop and apply methods/techniques in different aspects for clinical text understanding.

To answer the research questions discussed in Section 1.3, we propose concepts enhanced clinical document clustering, symptom/medication matching algorithms, graph regularized word embedding models, and their applications to clinical text understanding. The following is a summary of our contributions in the methods:

(1) Concept Enhanced Clinical Document Clustering Method.

We use Nonnegative Matrix Factorization (NMF) to integrate different features, such as words

and clinical concepts, for clinical document clustering. This provides a feasible way for us to visualize clinical documents at corpus level. Compared with general document clustering method, we discovered that extracted clinical concepts play an important role for clinical document clustering. We also use the method to analyze and visualize the risk factors for heart disease in the diabetic population. Our method integrates multiple risk factors with various attributes into uniform feature representations, and clusters patients' data from multiple aspects. This study explores new ways of visually interpreting risk factors for patients and assisting decision making for physicians.

(2) Symptom/Medication Relation Modeling and Recommendation.

Based on clinical concepts extracted from clinical notes, we build a symptom-medication (Symp-Med) graph to model symptom and medication relations in a corpus level. We develop two Symp-Med matching algorithms to predict and recommend medications for symptoms.

(3) Graph Regularized Word Embedding Models.

First, we propose graph-regularized word embedding models enhanced by KBs. Experiments on both general domain datasets and biomedical NLP tasks proof that Integrating extra knowledge can improve the performance of word embedding models.

Second, we apply the graph-regularized word embedding model and present a novel approach to automatically infer the most probable diagnosis from a given clinical narrative. Previous work on diagnosis inference from clinical narrative either formulating it as a medical literature retrieval task [25, 26] or solving it with multiclass algorithms in a supervised way [27]. We innovatively work on diagnoses inference from clinical narratives in an unsupervised way. Thus, we build baselines for this novel task.

## 1.5 Thesis Organization

The rest of this thesis is organized as following parts. Chapter 2 introduces the previous related work to our study. Chapter 3 and Chapter 4 address the question of *How to model the relationships from clinical notes at corpus level*. Chapter 3 presents our work on concept enhanced clinical document clustering for patient analytics. Chapter 4 presents our work on symptom/medication relation modeling from clinical notes. Chapter 5 and Chapter 6 address the question of *How to integrate*

*structured knowledge with unstructured text to improve results for Clinical NLP tasks?* Chapter 5 discusses our work of applying word embedding models to clinical NLP tasks. Chapter 6 explores the problem of diagnosis inference from clinical text. Finally, we conclude this thesis and introduce future directions in Chapter 7.

# Chapter 2: Related Work

## 2.1 Clinical Concept Extraction

### 2.1.1 Clinical Notes

A clinical note provides details about patient encounters and it's prepared by healthcare professional in unstructured text format. There are different types of clinical notes/reports generated for various purposes and from different patient visiting occasions, such as physician visit note, admission note, discharge summary, nursing progress notes, cardiac catheterization report, ECG report, radiology report, and echo reports[1]. In general, clinical note can be organized into four SOAP[2] sections [28] as follows:

- *Subjective*: patients verbally express symptoms and observations. Also details about medication history, family history, and etc.
- *Objective*: Observations include symptoms that can be measured in different ways, such as physical examination, test result, blood pressure, height, weight, and other vital signs.
- *Assessment*: a list of diagnoses regarding a patient's condition.
- *Plan*: follow-up directions for the patient, such as medications, treatment plan, and etc.

Even clinical notes have such loosely organized structure in general, important clinical concepts are distributed embedded in unstructured or semi-structured free text. Figure 2.1 is an example of clinical note. In this clinical note, patient's treatment history and plan are expressed in the free text of "*He was initially treated with antibiotic therapy. ... He was discharged home on Neupogen.*" Sophisticated clinical NLP tools are required to understand the narratives. "antibiotic therapy" needs to be identified as a "treatment" clinical concepts type, while "Neupogen" should be identified as a "medication".

[1] https://physionet.org/mimic2/mimic2_clinical_overview.shtml
[2] http://www.physiciansoapnotes.com/

```
ADMISSION DATE :
10/17/95
DISCHARGE DATE :
10/20/95
HISTORY OF PRESENT ILLNESS :
This is a 73-year-old man with squamous cell carcinoma of the lung , status post
lobectomy and resection of left cervical recurrence , admitted here with fever and
neutropenia .
Recently he had been receiving a combination of outpatient chemotherapy with the
CAMP Program .
Other medical problems include hypothyroidism , hypercholesterolemia , hypertension
and neuropathy from Taxol .
HOSPITAL COURSE :
He was started on Neupogen , 400 mcg. subq. q.d.
He was initially treated with antibiotic therapy .
Chest x-ray showed questionable nodule in the right lower lobe , reasonably stable .
Calcium 8.7 , bilirubin 0.3/1.3 , creatinine 1.1 , glucose 128 .
Hematocrit 24.6 .
WBC rose to 1.7 on 10/19 .
The patient had some diarrhea .
There was no diarrhea on 10/20 .
He was feeling well and afebrile .
The neutropenia resolved and he was felt to be in satisfactory condition on
discharge on 10/20/95 .
He was discharged home on Neupogen .
```

Figure 2.1 An Example of Clinical Notes (Data from Sun et al [2]).

### 2.1.2 Clinical Concept Types

Lots of efforts have been made in recent years to classify semantic types for clinical concepts. Terminology and ontology are built to describe concepts in the biomedic domain. For example, The Unified Medical System (UMLS) Metathesaurus [29] contains millions of biomedical and health related concepts. They are maintained by The National Library of Medicine (NLM). UMLS Metathesaurus defines clinical concept types in a very detailed level[3], such as:

- Finding

  Laboratory or Test Result

  Sign or Symptom

- Organism Attribute

  Clinical Attribute

- ...

[3]https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

There are other ontologies and classification systems developed for defining clinical concepts. For example, SNOMED CT [30] is a database contains terms and concepts for the coding of diagnosis and problem lists by clinicians. LOINC [31] is a database provides a universal code system for reporting laboratory and other clinical observations. RxNorm [32] provides clinical drug names and links its names to many of the drug vocabularies. ICD-10 (the 10th revision of the International Statistical Classification of Disease and Related Health Problem) [33] is a medical classification, which contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases[4]. These ontologies and classification systems are widely used to assist information extraction from biomedical text.

Most of clinical concept extraction tools do not focus in a very detailed level as described in the ontology and classification system. They refer clinical concepts in a more general level. Clinical concepts refer to names of *findings*, *treatment*, *test*, *disease*, *anatomy*, *substance*, *demographics*, and etc. For example, following clinical concepts can be extracted from example in Figure 2.1:

- **demographics:** 73-year-old, man, ...
- **test:** chest x-ray
- **findings:** fever, neutropenia, feeling well, ...
- **substance:** Neupogen, ...
- **others:** ...

### 2.1.3 Clinical Concept Extraction

Compared to concept extraction from free text in general domain, clinical concepts extraction needs to overcome the barriers of lack of annotated data, limited access to data, the variation of clinical text, and limited extra knowledge sources.

There are some existing systems and tools [16, 22, 23] for clinical concept extraction. MetaMap [34] is developed by NLM to extract Metathesaurus concepts from texts. It returns different semantic types presented in the text. Lots of clinical concept extraction tools are built based on MetaMap

[4] http://www.who.int/classifications/icd/en/

**Sentence Splitting:**
[This is a 73-year-old man with squamous cell carcinoma of the lung, status post lobectomy and resection of left cervical recurrence, admitted here with fever and neutropenia.][]…
**Tokenizing:**
[This | is | a | 73-year-old | man | with | squamous | cell | carcinoma | of | the | lung | , | status | post | lobectomy | and | resection | of | left | cervical | recurrence | , | admitted | here | with | fever | and | neutropenia | . |]…
**POS Tagging:**
This/DT is/VBZ a/DT 73-year-old/JJ man/NN with/IN squamous/JJ cell/NN carcinoma/NN of/IN the/DT lung/NN ,/, status/NN post/NN lobectomy/NN and/CC resection/NN of/IN left/JJ cervical/JJ recurrence/NN ,/, admitted/VBN here/RB with/IN fever/NN and/CC neutropenia/NN ./.
**NER:**
This is a [73-year-old]DEMOGRAPHIC [man]GENDER with [squamous cell carcinoma of the lung]DISEASE, [status post lobectomy]PROCEDURE and [resection of left cervical recurrence]PROCEDURE, admitted here with [fever]SYMPTOM and [neutropenia]SYMPTOM.

Figure 2.2 Subtasks in Clinical Concept Extraction.

[35]. cTAKES (Mayo clinic's clinical Text Analysis and Knowledge Extraction System [16]) is an open source NLP system for clinical concept extraction. HITEx (Health Information Text Extraction [36]) is an open-source NLP system for extracting clinical concepts like diagnosis, discharge medications, smoking status, and etc. MedEx [15, 37] is an open source system processing clinical text and extracting medication names and signature information, such as drug dose, frequency, route, and duration. It designed for medication information extraction and reported a 93.2% F-measure on identifying drug names. Reference [38] proposes a method to identify medical concepts from the SNOMED Clinical Terminology in free texts. Another linguistic approach for identification of medication names and related information in clinical narratives uses negation maker to exclude negation medication information [39]. NegEx [40] is a tool for determining findings and diseases from the clinical text are negated or not. More details about existing clinical concept extraction systems can refer to systematic reviews in [41, 42].

These systems usually contain basic NLP modules for clinical text processing, such as sentence splitting, tokenization, part-of-speech (POS) tagging, Name Entity Recognition(NER), and etc, as displayed in Figure 2.2. These clinical concept extraction systems are highly dependent on ontologies and dictionaries (discussed in section 2.1.2) from biomedical domain. These domain ontologies include UMLS Metathesaurus [29], ICD-10 classification [43], RxNorm [32], SNOMED-CT [30], and etc.

## 2.2 Clinical Document Clustering

Document clustering techniques provide an efficient way of navigating and summarizing documents into a small number of meaningful clusters. They have received lots of attentions in recent years [44, 45]. Nonnegative Matrix Factorization (NMF) is a clustering algorithm to factorize a matrix $V$ into two matrices $W$ and $H$, all three matrices have no negative elements. NMF has been widely applied to document clustering [46, 47]. Akata et al [48] extended NMF towards joint NMF, which can jointly analyze different types of features for multi-view learning. Instead of fixing a common clustering solution for each view, Liu et al [49] further formulated the process by finding the nearest consensus for each view. Multi-view NMF can integrate various sources of data and yield a better clustering result [50]. In our study (described in 3.5), we apply multi-view NMF to integrate features of symptom concept, medication concept, and word from the clinical document for clustering.

For clinical document clustering [51], Saad et al [52] investigated clinical documents clustering for grouping clinical documents into meaningful clusters and discovering patterns and important features. Patterson et al [53] clustered a data set consisting of 17 clinical note types using an unsupervised clustering algorithm and demonstrated different clinical domains use different lexical and semantic patterns. Doing-Harris et al [54] identified medical specialty across institution by comparing linguistic features of clinical notes from different institutions using document clustering techniques. Han et al [55] employed latent semantic indexing to cluster clinical notes and found that latent semantic indexing was an effective method for measuring the similarity of clinical notes. Zhang et al [56] evaluated nine semantic similarity measures of ontology-based terms for medical document clustering. Documents clustering provides an efficient way for physicians and patients to understand patterns inside and among clinical documents.

## 2.3 Clinical Relation Extraction

For a given pair of entities, relation extraction is defined as classifying the relation between this pair of entities into one of the predefined relation types or no relation [57]. General relation extraction methods and models can be classified as unsupervised, semi-supervised, distant-supervised, open IE,

and etc. The commonly used features for relation extraction include lexical, syntactic, semantic, contextual, and etc. Some recent work [58] learned relation extraction model jointly from KBs and text. Embedding representation for words and entities/relations in KBs are explored to facilitate the relation extraction tasks [59, 60].

### 2.3.1 Clinical Relations

Relation extraction can facilitate clinical decision making. The UMLS [61] Metathesaurus contains a large amount of manually extracted relations for UMLS concepts. However, these relations stored in such KBs is far from complete, and the knowledge is changing extremely fast. Thus, lots of research work explore to automatically extract clinical relations from text corpus for complementing existing manually created relational KBs. Wang and Fan [62] presented a manifold model to extract medical relations from sentences corpus. Hassan et al [63] focused on methods to automatically extract disease-symptom relationships from text. Disease-symptom relationship is an important type of relationship. Rosario and Hearst [64] focused on extracting relationship between the entities "treatment" and "disease" from bioscience text. Most of the current clinical relation extraction research work focused on one particular type of relation and evaluated based on a limited number of manually created datasets. Lally et al [3] created an Emerald model to systematic summarizing different types of relationships in the medical domain, as shown in Figure 2.3.

The Emerald model provide a comprehensive overview of clinical relationship types. Some commonly used clinical relations are "*findingOf*" relation type between clinical types "*finding*" and "*disease*", "*treats, prevents*" relation type between "*disease*" and "*treatment*", "*diagnoses*" relation type between "*test*" and "*disease*", and etc.

### 2.3.2 Symptom/Medication Relation Extraction

Clinical symptoms are important for patients to control the exacerbation of diseases [65]. Reference [28] proposes a framework for modeling and mining symptom relationships from clinical notes. The relationships between symptoms and medication for one particular disease (such as asthma [66, 67], cancer [68]) have been studied with case study methods and statistical methods. A symptom-

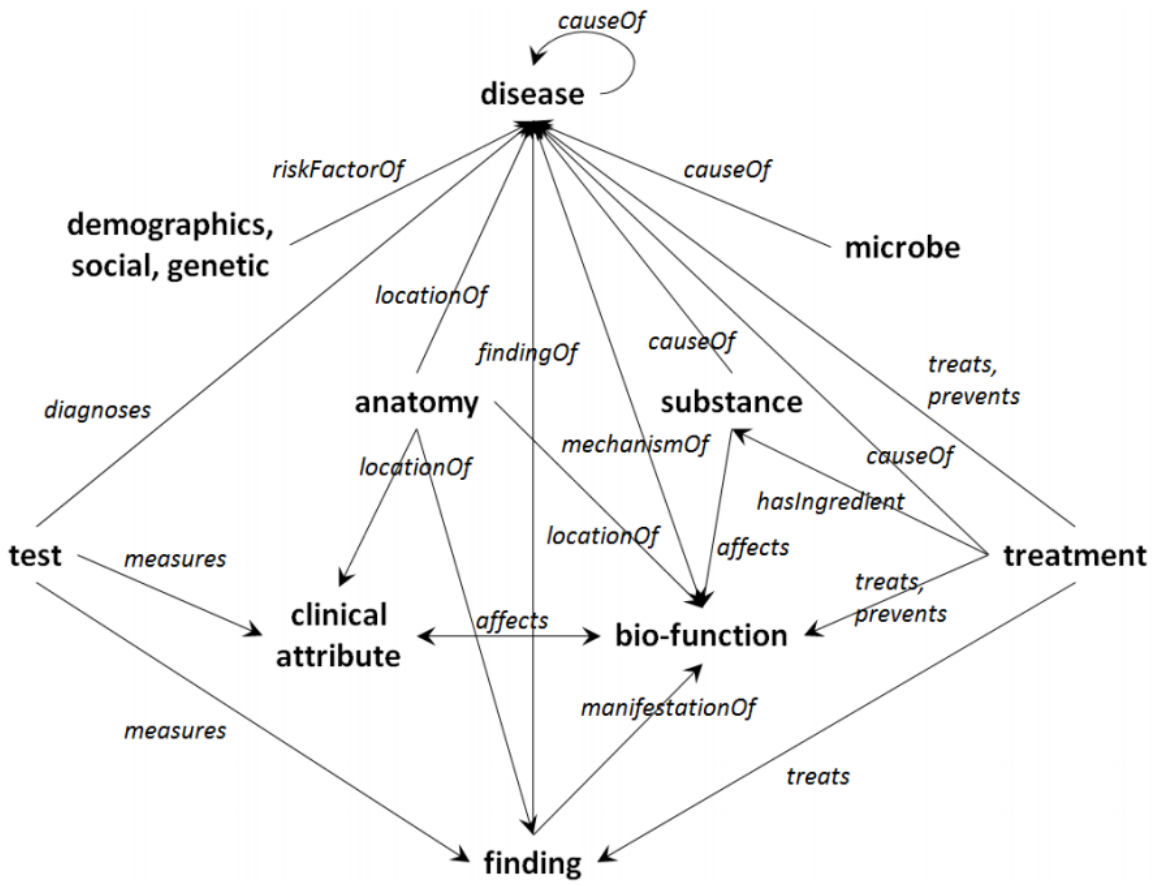



Figure 2.3 The Emerald [3].

medication score is used as an instrument to evaluate the disease severity by recording symptoms and rescue medication [69]. Currently, there is little research work on extracting symptom and medication concepts from clinical notes for medication error detection and surveillance. In our study (in Section 4), We propose a Symptom-Medication matching framework to model symptom and medication relationships from clinical notes. After extracting symptom and medication concepts, we construct a weighted bipartite graph to represent the relationships between the two groups of concepts. The key is to efficiently answer user's symptom-medication queries using the graph.

Bipartite graph is a commonly used to represent relationships among concepts extracted from texts. SympGraph [28] uses the bipartite graph to represent symptom information from clinical notes. A bipartite graph contains two groups of vertices connected between groups and no edge among the vertices in the same group. Maximum matching is an important problem for bipartite

graph [70]. Reference [71] develops neighborhood formation and anomaly detection algorithms for the bipartite graph. The neighborhood formation algorithm is to find similar vertices inside a group, which can be used for symptom expansion and medication expansion.

## 2.4 Word Embedding Models

There is a growing trend of applying embedding representation to improve feature representations for information extraction tasks in the biomedical domain [72, 73]. Prior work indicated that using word embedding can significantly improve the performance of concept extraction tasks [74, 75]. For relation extraction, a recent work [76] evaluated word embedding on medical corpora, it showed there are necessities to have more in-depth work on applying word embedding in the medical domain.

### 2.4.1 Embedding Representation

One basic feature representation method has been applied to a vast majority of NLP tasks is called one-hot representation [77]. The one-hot representation preserves the original form of the feature in a vector. For example, for given a set of word features $\{cat, dog, sheep, cow, \dots\}$, a single word *cat* can be represented in the feature vector space with one 1 and a lot of zeros:$[1, 0, 0, 0, \dots]$. The one-hot representation has two limitations: First, the similarity between features cannot be captured. For example, a single word *cat* is represented as $[1, 0, 0, 0, \dots]$, and word *dog* is represented as $[0, 1, 0, 0, \dots]$. The similarity between these two feature vectors is 0. But in reality, there is some level of semantic similarity between these two words, the one-hot representation fails to capture such similarity. Second, when the features set is large, the feature vector has a high dimension, the computation cost will be expensive.

To address the limitations of one-hot representation, distributed representation [78] has been studied for a long time. Distributed word representations were introduced by [78], and have been successfully applied to many NLP problems through neural network based language models [79, 80, 81, 82, 83, 84]. Later, Mikolov et al [85, 86] proposed two-word embedding methods: the continuous bag-of-words model (CBOW) and the continuous skip-gram model. The word embedding model is to learn distributed representations of words from a large amount of unlabeled text data, a word is

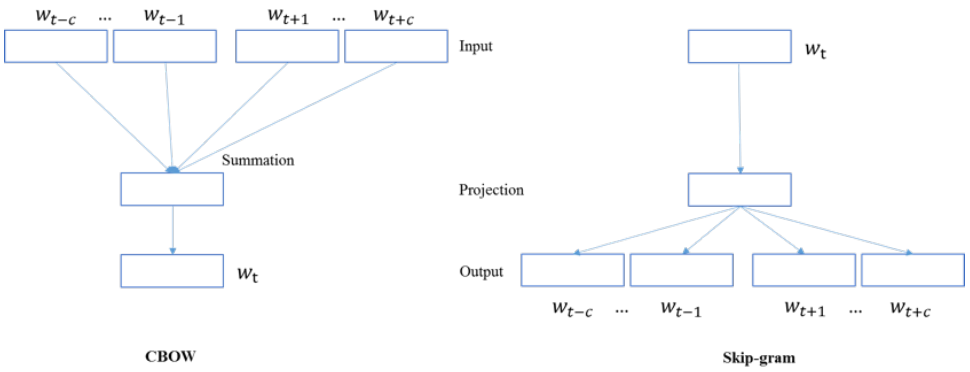



Figure 2.4 The CBOW Architecture and Skip-gram Architecture.

represented as a dense and low-dimensional vector. The semantically similar words will be transferred into similar vector representations. They developed two training methods, hierarchical softmax and negative sampling, to train both CBOW and Skip-gram models. Experimental results showed that vectors learned from these two models yielded state-of-the-art performance on word similarity tasks. Examples of further illustrations of the CBOW and Skip-gram models can be found in [87, 88, 89].

### 2.4.2 CBOW and Skip-gram Models

Word embedding models learn distributed representations of words from a large amount of unlabeled text data. Each word is represented as a dense and low-dimensional vector, and semantically similar words are transformed into similar vector representations.

Both CBOW and Skip-gram models are three-layer neural networks, containing input, projection, and output layers, as displayed in Figure 2.4. The CBOW model learns word embedding by using context words to predict the center word $w_t$, where the context words refer to the neighbouring words within a window size $c$ near the centre word in a sentence. Given a sequence of training words $w_1, w_2, \ldots, w_T$, the CBOW model has the following objective function:

$$J_1 = \max \frac{1}{T} \sum_{t=1}^{T} \log p(w_t | w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}) \tag{2.1}$$

The Skip-gram model predicts surrounding words $w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}$ given the current

centre word $w_t$. This model has the following objective function:

$$J_2 = \max \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) \tag{2.2}$$

The probability $p(w_t|w_{t+j})$ is calculated using a softmax function:

$$p(w_t|w_{t+j}) = \frac{\exp(v_t^{'T} v_{t+j})}{\sum_{n=1}^{N} \exp\left(v_n^{'T} v_{t+j}\right)} \tag{2.3}$$

$v_n$ and $v_n^{'}$ are the input and the output representation vectors of word $w_n$. $N$ is the total vocabulary size. The representation vectors $v_n$ are between the input layer and projection layer, and $v_n^{'}$ are between projection layer and the output layer.

In the CBOW model, the projection layer $h$ is the average value of input representation of context words.

$$h = \frac{1}{2c} \sum_{-c \leq j \leq c, j \neq 0} v_{t+j} \tag{2.4}$$

In the Skip-gram model, the projection layer $h$ is the same as the input representation of word $w_t$, which is $v_t$.

### 2.4.3 KB Enhanced Word Embedding Models

Since word embedding models are trained through an unsupervised manner, the learned distributed representations may contain some noises. Therefore, recent studies start to explore incorporating different types of resources as auxiliary supervision to improve the performances of distributed representations [90, 91, 92, 93, 94, 95, 96]. For example, Zhou et al. [97] proposed to apply metadata of category information from community question answering to enhance learning word embedding representation. Experimental results showed that extra knowledge can improve Skip-gram model on question retrieval task.

Of all types of resources, KBs have attracted a lot of attentions and have proven valuable for improving the performances of word embedding models. Bian et al. (2014) [98] explored three

types of knowledge as additional input information to serve as auxiliary supervision in CBOW model. The three types of knowledge include semantic knowledge, morphological knowledge, and syntactic knowledge. The knowledge is acquired from four resources: Morfessor [99], Longman Dictionaries[5], WordNet [100], and Freebase [101]. Experimental results on analogical reasoning task, word similarity task, and word completion task demonstrated that these extra knowledge resources can enhance the performances of the CBOW model. Xu et al. (2014) [102] introduced a new RC-NET framework to leverage both relational and categorical knowledge by integrating them as two separate regularization functions into the original optimization problem in order to enhance Skip-gram model. They applied knowledge from WordRep (Gao et al., 2014) and Freebase. Experimental results on analogical reasoning task, word similarity task, and topic prediction task showed that the quality of distributed representations is improved. Wang et al.(2014) [103] examined the relations between entities from large-scale knowledge graph and proposed a method that jointly embeds entities and words into the same continuous vector space. They used Freebase as their knowledge source and compared their method with Skip-gram model on the analogical reasoning task. Liu et al.(2015) [104] proposed to incorporate semantic knowledge into Skip-gram model. Semantic knowledge is presented as ordinal ranking inequalities to formulate the word embedding learning problem as a constrained optimization problem. They obtained semantic knowledge from WordNet. Experimental results on word similarity task, sentence completion task, name entity recognition task, and TOEFL synonym selection showed that distributed representations can be improved by incorporating semantic knowledge.

In our study (in section 5), we leverage structured KBs to enhance state-of-the-art word embedding models. In contrast to the aforementioned studies on enhancing word embedding representation with extra knowledge from KBs, we design a graph regularized framework to improve both CBOW model and Skip-gram model. Compared with Xu et al.(2014) [102], we have different regularization framework and have it tested on both CBOW and Skip-gram. Our work is motivated by the research work of using graph regularization to improve data representation [105]. Ordinary Non-

[5]http://www.longmandictionariesonline.com/

negative Matrix Factorization (NMF) only considered the Euclidean structure of data, but ignores their semantic relationships. Instead, with a graph regularized NMF (GNMF), which can uncover the hidden semantics and respect the intrinsic geometric structure. GNMF achieved better results. Compared to other extra knowledge resources, a structured KB can assist to encode semantic information with the graph structures. Thus, our work is to take advantage of such structure to improve word embedding models performance, in both CBOW and Skip-gram.

## 2.5 Clinical Diagnosis Inference

Clinical diagnostic inferencing is a challenging task. For example, given a clinical case (past medical history, signs and symptoms etc.), the clinician administers appropriate medical tests or procedures, infers the accurate diagnosis, and prescribes the best possible treatment plan based on his/her experience or up-to-date knowledge/evidence obtained through substantial research on relevant external resources. Recent works on diagnostic inferencing mostly use recurrent neural networks (RNNs) by utilizing structured clinical data e.g. physiological signals, vital signs, lab tests, and other variables [106, 107]. Clinical diagnostic inferencing can be regarded as one type of clinical questions, that is "*what is the patient's diagnosis?*" Related research [108] become more and more popular due to the increasing availability of clinical datasets for public recently.

For general Question Answering (QA), there are two major parts: first, question interpretation. Understanding the question correctly is critical to QA. Semantic parsing [109, 110] has been applied to question interpretation. The second part is question-answer matching [111]. Question interpretation will convert the question to the database query. Some large-scale structured KBs (such as Freebase) are frequently used to match question to answers [112, 113]. One recent research trend [97, 114, 115, 116] is to apply word embedding for question-answer matching.

Regardless the rapid development in general QA domain, clinical QA, especially scenario-based question analysis [117, 3], still requires lots of input from the domain expert and a variety of sources (such as knowledge encyclopedia and domain-specific knowledge bases), and some of these resources are not easily accessed by researchers. A recent work [118] proposed a subgraph embedding model to Question answering. The subgraph embedding model learns low-dimensional embedding for words

and knowledge base constituents and uses the representations to score question against candidate answers. The subgraph embedding makes the assumption that all potential answers are entities in the KB and question word sequence contain one identified KB entity. There are two limitations when applied to CQA: (1) The potential answers may not be an entity in KB; (2) The keywords extracted from clinical question narratives could be more than one, and they may not find the identified form in KB.

The Text Retrieval Conference (TREC 2014 [119], TREC 2015 [120], and TREC 2016 [6]) released a Clinical Decision Support (CDS) task. The task requires to retrieval relevant biomedical articles for clinical reports to answer three types of generic clinical questions: Diagnosis ("*what is the patient's diagnosis?*"), Tests ("*what tests should the patient receive?*"), and Treatment ("*How should the patient be treated?*"). The accurate identification of diagnosis is proved to be helpful to biomedical articles retrieving and the other two types of questions answering [121]. The development of large-scale structured KB in medical domain (such as UMLS [29]) and a lot of available data sources (such as MIMICII database [122], EMR, and etc.) promote the research work on clinical diagnosis inference. However, due to the difficult to interpret clinical narratives and the lack of complete domain knowledge, the clinical diagnosis inference is still a challenging problem.

## 2.6 Available Sources for Clinical Text Research

In this section, we summarize different types of available data sources for clinical text research.

### 2.6.1 Knowledge Bases (KBs)

One frequently mentioned knowledge resource for clinical NLP tasks is structured Knowledge Base (KB). We have witnessed a quick development of KBs in past years. KBs store structured information about entity types and relation triples. A triple is represented as $\langle subject, predicate, object \rangle$. Many large-scale KBs of general or specific domains have been constructed, such as WordNet [100], Yago [123], Freebase[101], DBpedia [124], and NELL [125], UMLS [29]. KBs are useful resources and powerful tools for supporting NLP tasks such as relation extraction [126, 127] and question answering [118].

[6] http://trec-cds.appspot.com/2016.html

For structured KBs, UMLS Metathesaurus and Freebase provide information about semantic relation triples that are related with biomedical related concepts. The UMLS MRREL table defines the relationships between UMLS concepts. One example relation triple in the table is $\langle concept : Giardiasis; relation : may\_be\_treated\_by; concept : Furazolidone \rangle$.

Freebase: Freebase [3] is a knowledge base contain many triples, such as ⟨ subject; predicate; object⟩. There are triples from freebase that are related with medicine relation types. One example relation triple in freebase is $\langle Giardiasis; medicine.disease.symptoms; Flatulence \rangle$.

### 2.6.2 Shared Tasks and Datasets

Shared tasks in the biomedical domain have existed for over two decades [128]. There are different types of shared tasks in clinical NLP area [129]. They provide de-identified clinical notes for participants, which is one most common way for researchers to get access to clinical data.

i2b2 (Informatics for Integrating Biology and the Bedside)[7] is an NIH-funded national center for biomedical computing based at partner healthcare system. They provide clinical data for researchers through the i2b2 NLP shared tasks. The i2b2 NLP shared tasks started from 2006. The first challenges is de-identification [130] and smoking status classification [131]. The smoking status classification provided a corpus of 502 discharge summaries [13]. Other i2b2 challenges are summarized as follows:

- **2008 obesity challenge** [132]. The challenge consisted of 1237 discharge summaries of patients who were overweight or diabetic and had been hospitalized for obesity or diabetes. The task targeted on identifying obesity and its comorbidities.
- **2009 Medication Challenge** [133, 17]. This medication challenge contained a total of 1243 deidentified discharge summaries, which were used for extracting medications and associated information.
- **2010 Relations Challenge** [129]. The challenge contained a total of 394 training reports, 477 test reports, and 877 unannotated reports of discharge summaries, which were used for

[7] https://www.i2b2.org/

identifying concepts, assertions, and relations.

- **2011 Coreference Challenge** [134]. The challenge provided training set and the test set from four hospitals. The training set includes 492 labeled discharge summaries. The test set consists of 322 discharge summaries [135].

- **2012 Temporal Relations Challenge** [136, 2]. This challenge provided 310 de-identified discharge summaries, with annotations of clinical events, temporal expressions, and temporal relations.

- **2014 De-identification and Heart Disease Risk Factors Challenge** [137, 137]. The challenge provided a set of 1304 longitudinal de-identified medical records describing 296 patients.

Each of these shared tasks included a corpus of clinical narratives, and these corpora are available from http://i2b2.org/NLP/DataSets with a data use agreement.

THYME corpus [138] is a collection of over 1,200 de-identified notes from the Mayo Clinic, representing patients from the oncology department, specifically those with brain or colon cancer. The corpus was used for SemEval 2015 task [139]. Deleger et al., [140] created a corpus of 3,503 de-identified medical records of 22 different types, including discharge summaries, progress notes, and referrals. Text REtrieval Conference (TREC) also provides medical records corpus. TREC 2012 Medical Records Track [141] contained 93,551 reports mapped into 17,264 visits.

TREC CDS (Clinical Decision Support) track[8] investigated techniques for linking medical records to information relevant to patient care. TREC CDS 2014 [19], 2015 [120], and 2016[9] all provided actual medical records, which are well-formed narratives summarizing the portions of patients' medical record that are pertinent to the case.

### 2.6.3 Open Access Text Sources

For unstructured text sources, there are different types of resources can be used:

[8] http://trec-cds.appspot.com/
[9] http://trec-cds.appspot.com/2016.html

**MIMIC-III** [142] (Medical Information Mart for Intensive Care) is a large database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. MIMIC-III contains data associated with 53,423 distinct hospital admissions for adult patients (aged 16 years or above) admitted to critical care units between 2001 and 2012. MIMIC-III is one of the most widely-used collections of clinical notes.

**PMC**[10] (PubMed Central) is an online digital database of freely available full-text biomedical literature. It has been broadly applied to clinical NLP tasks [143, 144].

**Wikipedia**[11] has a large collection of pages. It has clinical diseases and medicine related pages under clinical medicine category[12].

**DailyMed**[13] contains documents describing drugs.

**WebMD**[14] contains documents describing drugs. Each document contains the same 7 sections, such as Uses, Side Effects, Precautions, etc.

**MayoClinic**[15] pages include Symptoms, Causes, Risk Factors, Treatments and Drugs, Prevention, etc.

These open access text sources are widely utilized for supporting clinical NLP tasks [145, 146, 147].

[10] https://www.ncbi.nlm.nih.gov/pmc/
[11] https://www.wikipedia.org/
[12] https://en.wikipedia.org/wiki/Category:Clinical_medicine
[13] dailymed.nlm.nih.gov
[14] www.webmd.com
[15] www.mayoclinic.org

# Chapter 3: Clinical Document Clustering

Clinical documents are rich free-text containing valuable information. In this chapter, we explore to use document clustering methods to analyze the intrinsic patterns in clinical documents corpus. Concept extraction is the very first step for clinical text understanding, thus we build an integrating system for extracting medication names and symptom names from clinical notes. Based on concept extraction, we further explore different ways of using document clustering techniques to cluster clinical documents into meaningful clusters and analyze latent patterns from clinical documents [148].

## 3.1 Motivation

Clinical notes contain valuable information about patients, such as medication conditions (diseases, injuries, medical symptoms, and etc.) and responses (diagnoses, procedures, and drugs) [149]. These underutilized resources have a huge potential to improve health care. Different types of valuable information extracted from clinical notes can be used to build profiles for individual patients [150], discover disease correlations [151] enhance patient care [152], and etc.

Symptoms and medications are two important types of information that can be obtained from clinical notes. Symptom-related information such as diseases, syndromes, signs, diagnose etc., can be used to analyze diseases for patients. In addition, valuable medication information is commonly embedded in unstructured text narratives spanning multiple sections in clinical documents [153]. Medication information from clinical notes is often expressed with medication names and other signature information about drug administration, such as dosage, route, frequency, and duration. Thus symptom information and medication information extraction for clinical notes usually need sophisticated clinical language processing methods [10]. We want to explore efficient ways to extract symptoms and medications names from clinical notes.

Recently, large volumes of clinical documents are generated by electronic health record systems

**PRINCIPAL DIAGNOSIS:** *Cellulitis*.
**LIST OF PROBLEMS/DIAGNOSES:** 1. *Inflammatory breast cancer*. 2. *Type II diabetes*. 3. *Hypertension*. 4. *Hypercholesterolemia*. 5. *CHF* with preserved systolic function. 6. *Obstructive sleep apnea*. 7. *Asthma*. 8. *Spinal stenosis* with herniated discs.
**MEDICINES:** 1. *Lantus* 40 units nightly. 2. *Aspirin* 81 mg daily. 3. *Lipitor* 40 mg daily. 4. *Zestril* 2.5 mg daily. 5. *Cardizem ER* 240 mg daily. 6. *Lasix* 20 mg daily. 7. *Procrit* 40000 units weekly. 8. *Pamidronate*. 9. *Dexamethasone* with chemotherapy. 10. AC chemo. 11. *Neulasta*. 12. *Ativan p.r.n*. 13. *Multivitamin*. 14. *Iron sulfate*. 15. *Isosorbide dinitrate* 10 mg t.i.d. 11. *Allegra* 60 , 000 mg b.i.d.

Figure 3.1 A Clinical Note Example with Several Selected Sections.

[154, 155]. Different types of clinical notes are generated for various of purposes and from different occasions. For example, patient admission note, discharge note, laboratory report, and etc. Among them, physician's visit notes are one of the most important notes for patients. They are generally organized into four SOAP sections [28]. SOAP [1] standards for subjective, Objective, Assessment and Plan. Even these clinical notes have such loosely organized structure, important medical concepts still exist in the unstructured or semi-structured text. Due to the individual diversity of clinical narratives, it is a challenging problem to discover the underlying patterns from a corpus of clinical documents.

## 3.2 Clincal Notes Analysis

Clinical notes are an important format of patient records, and most of the clinical notes are in unstructured free-text format. An example of clinical note with a few selected sections is displayed in Figure 3.1. There are three sections included in this clinical note example: *Principal Diagnosis*, *List of Problems/Diagnoses*, and *Medicines*.

Symptom and medication names are mentioned in this clinical note example. As shown in Figure 3.1, these highlighted names are embedded in multiple sections of unstructured/semi-structured text. The symptom and medication names are important concepts delivering information about patients, disease progression. They are critical for physicians and patients to understand this clinical

[1] http://www.physiciansoapnotes.com/

note.

We conduct statistical analysis on one of our experiment dataset (Details in Section 3.5.1). The most frequent sections in clinical notes contain medication/symptom names are displayed in Table 3.1. *Discharge Medications*, *History of Present Illness*, *Hospital Course*, *Brief Resume of Hospital Course*, *Hospital Course By System*, and *Hospital Course By Problem* are six most frequent sections contain both symptom names and medication names.

Table 3.1 Most Frequent Clinical Notes Sections with Medication/Symptom Names

| **Most Frequent Sections with Symptom Names** | **Most Frequent Sections with Medication Names** |
|---|---|
| Amit Diagnosis | Discharge Medications |
| History Of Present Illness | Hospital Course |
| Hospital Course | History Of Present Illness |
| Past Medical History | Potentially Serious Interaction |
| Brief Resume Of Hospitlal Course | Medications On Admission |
| Discharge Medications | Brief Resume Of Hospitlal Course |
| HPI | Medications |
| Physical Examination | Medications On Discharge |
| Hospital Course By System | Hospital Course By System |
| Hospital Course By Problem | Hospital Course By Problem |

## 3.3 Concepts Extraction Framework

In this section, we present a clinical concepts extraction framework. An overview of extracting symptoms and medications from clinical notes is showed in Figure 3.2. We build the framework to extract the symptom names such as "hypertension" and medication names such as "Isordil, Cardizem" from the clinical texts "He was kept off aspirin given his GI bleeding. The patient also has hypertension and was on Isordil and Cardizem for that." The overall system contains five parts: word/sentence annotator; section annotator; negation annotator; symptom name annotator; and medication name annotator.

First, we process clinical notes to identify words and sentences from clinical notes using Stanford CoreNLP Tool[2]. During the pre-processing, we use section annotator to identify different sections for each clinical note. The section annotator depends on the section header information from clinical

[2]http://nlp.stanford.edu/downloads/

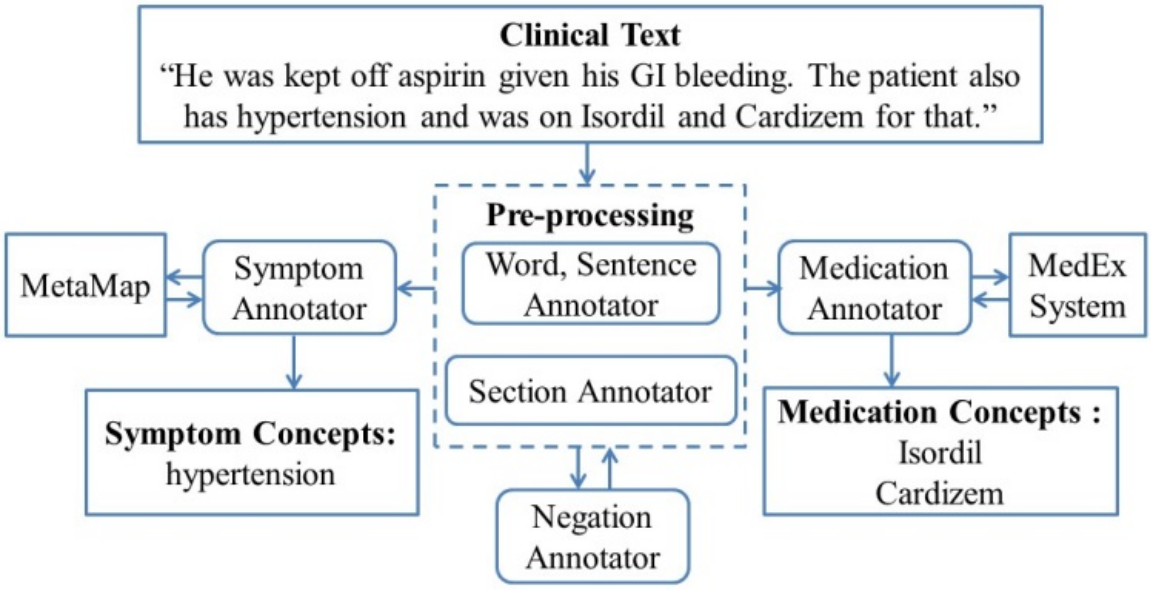



Figure 3.2 An overview of symptom/medical term extraction from Clinical Notes.

notes. Negation sections, such as "ALLERGIES" or "Family History", are excluded. For example, "She is allergic to MORPHINE" from the section "ALLERGIES", the medication name "MORPHINE" is a negation medication name, so we exclude it.

We also use negation annotator to remove negation symptom and medication names. An example is that "The patient was told to avoid taking aspirin or any other NSAIDs given his GI bleed", we remove "aspirin" and "NSAIDs" because of the pre-negation words avoid. Pre-negation and post-negation are defined in Negation maker (i.e. NegEx[3]). Pre-negation is negation words like *avoid*, *deny*, *cannot*, *without*, and so on. Post-negation is negation words like *free*, *was ruled out*, and so on.

After pre-process, we use symptom annotator based on the MetaMap [21] to extract symptom names from clinical notes. Meanwhile, we use medication annotator based on MedEx System [15] to extract medication names from clinical notes.

We use MetaMap to extract symptom names from clinical notes. MetaMap[4] is a program that maps biomedical texts to concepts in the UMLS Meta-thesaurus [21, 34]. Since Metamap returns all

[3]http://www.dbmi.pitt.edu/chapman/NegEx.html
[4]http://nls3.nlm.nih.gov

types of concepts, we only keep these concepts related to symptom names, such as concept labeled as "sosy", which represents "sign and symptom". The related types of concepts include: {*sosy*, *dsyn*, *neop*, *fngs*, *bact*, *virs*, *cgab*, *acab*, *lbtr*, *inpo*, *mobd*, *comd*, *anab*}, see [28] in detail.

We use MedEx system to extract medication names from clinical notes. The MedEx system is a natural language processing system to extract medication information from clinical notes [15].

In clinical notes, medication data are often expressed in medication names and signature information about drug administration. The MedEx system extracts multiple semantic categories of medication findings from clinical notes, such as DrugName, Strength, Route, Frequency, Form, Dose Amount, IntakeTime, Duration, Dispense Amount, Refill, and Necessity. Here we use the DrugName as medication name.

## 3.4 Document Clustering Methods

### 3.4.1 Nonnegative Matrix Factorization (NMF)

Nonnegative Matrix Factorization (NMF) is a useful tool for the decomposition of multivariate data [156, 157]. The basic idea for NMF is to factorize a $n \times m$ matrix $A$ into a nonnegative $n \times k$ matrix $W$ and a nonnegative $k \times m$ matrix $H$ to approximate:

$$A \approx W \times H \tag{3.1}$$

$W$ and $H$ are two lower dimensional non-negative matrices. The value $k$ in NMF can be explained as a number of "meaningful" clusters. The choice of $k$ value is important. In reference [158], they developed approach to decide $k$ based on a consensus matrix and cophenetic correlation coefficient.

As discussed in [105], there are two commonly used cost functions can be applied to the approximation. The first one is using the square of Euclidean distance:

$$C_1 = \|A - WH\|^2 \tag{3.2}$$

The cost function can be minimized by applying the update rules as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \tag{3.3}$$

$$W_{ia} \leftarrow W_{ia} \frac{(AH^T)_{ia}}{(WHH^T)_{ia}} \tag{3.4}$$

The second one is using the "divergence":

$$C_2 = D(A||WH) \tag{3.5}$$

The cost function can be minimized by applying the update rules as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{\sum_i W_{ia} A_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \tag{3.6}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} A_{i\mu}/(WH)_{i\mu}}{\sum_v H_{av}} \tag{3.7}$$

### 3.4.2 Multi-View NMF

NMF has been extended to multi-view learning. Multi-view learning aims to identify latent components in different sub-matrices in a simultaneous manner. These sub-matrices can represent different features spaces. Akata et al [48] extends the basic NMF to a convex combination of $p$ different views as following optimization problem:

$$\min_{W^i, H \geq 0} \sum_{i=1}^{p} \lambda_i \left\| A^i - W^i H \right\|^2, \tag{3.8}$$

$$\sum_{i=1}^{p} \lambda_i = 1, \lambda_i \geq 0 \tag{3.9}$$

Due to constraint that matrix $H$ is fixed among multiple views, Liu et al [49] further extend to solving the following optimization problem:

$$\min_{W^i,H^i,H^*\geq 0} \sum_{i=1}^{p} \left\|A^i - W^i H^i\right\|^2 + \sum_{i=1}^{p} \lambda_i \left\|H^i - H^*\right\|^2 \tag{3.10}$$

This problem attempts to optimize $A^i \approx W^i H^i$ for each view $i$, and keep constraining each $H^i$ will be similar.

## 3.5 Clinical Concepts Enhanced Document Clustering

In this section, we apply NMF and multi-view NMF to cluster clinical notes into meaningful clusters based on sample-feature matrices. Our experimental results show that multi-view NMF is a preferable method for clinical document clustering. Moreover, we find that using extracted medication/symptom names to cluster clinical documents outperforms just using words.

### 3.5.1 Datasets

We conduct experiments on two datasets, the two datasets are acquired from i2b2 workshop on NLP challenges [5] at two different years: 2009 clinical notes dataset [17] and 2014 clinical notes dataset [159, 137].

**Datasets Description**

2009 dataset contains 1249 clinical notes in total. After pre-processing, 1239 clinical notes remain. One clinical note example is displayed in Figure 3.1.

2014 clinical notes dataset contains 1304 records from 296 patients. Each patient has about 3-5 records. Compared with 2009 clinical notes dataset, this dataset was applied for the risk factor identification for heart disease track. All the risk factors are annotated in these records. We classify these risk factors into symptom names or medication names. The original records have standard to indicate three types of patients. The first type is patients who develop Coronary Artery Disease (CAD), the second type is patients who have CAD in their first records, and the third type is patients

[5] https://www.i2b2.org/NLP/

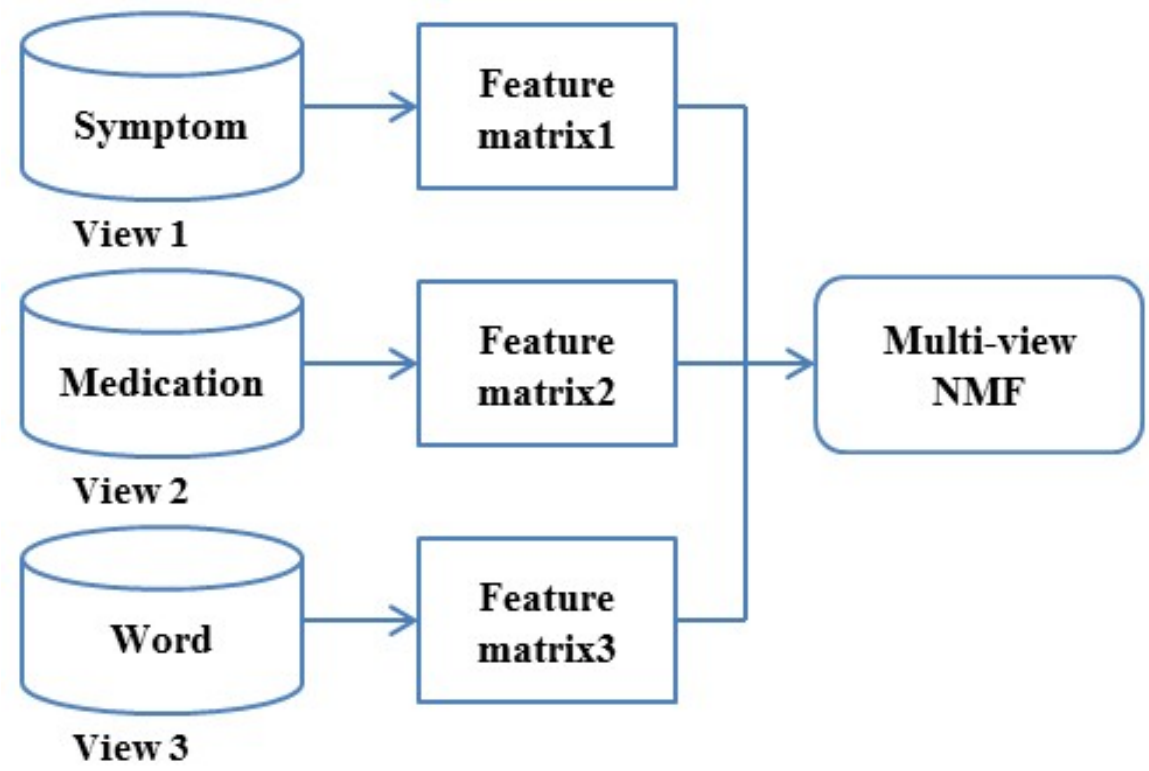



Figure 3.3 The Framework of Applying Multi-view NMF.

never develop CAD. We use this as standard to evaluate the cluster performance from multi-view NMF.

**Preprocessing**

We preprocess the dataset to generate the sample-feature matrices as shown in Figure 3.3. For 2009 dataset, we process each clinical record as a sample. While for 2014 dataset, each patient is processed as a sample. Each sample can be represented from three views: symptom names, medication names, and words. For the words set, we remove common stop words and clean the data. We generate features from these three views using word count or Term Frequency-Inverse Document Frequency (TF-IDF).

After preprocessing, we get these sample-feature matrices. The matrices' attributes are presented in Table 3.2.

For 2009 clinical notes dataset, the total number of symptom features is 2294, medication features

Table 3.2 Sample-Feature Matrices Size

| #Size | 2009 Clinical Notes Dataset | 2014 Clinical Notes Dataset |
|---|---|---|
| Samples | 1239 | 296 |
| Symptom Features | 2294 | 21 |
| Medication Features | 1029 | 18 |
| Unique Words | - | 17492 |

are 1029 correspondingly. For 2014 clinical notes dataset, medication features are 21, medication features are 18, and words feature are 17492.

### 3.5.2 Evaluation Metrics

For 2009 clinical notes dataset, since we don't have standard to evaluate the clustering result. We present and analyze the major features standout from each component factorized.

For 2014 clinical notes dataset, we use accuracy and normalized mutual information (NMI) as evaluation metrics [160].

Accuracy represents the number of correctly classified compared with known class labels. The higher accuracy means better performance.

NMI measures the clustering performance, the higher the better.

$$NMI = \frac{\sum_{h,l} n_{h,l} \log \frac{n_{h,l}}{n_h, n_l}}{\sqrt{\prod_{i=h,i} n_i \log \frac{n_i}{n}}} \tag{3.11}$$

Where $n$ represents the total number of documents, $n_h$ is the number of document in standard class $h$, $n_l$ is the number of documents in predicted cluster l, and $n_{h,l}$ is the number of documents in both clusters $h$ and $l$.

### 3.5.3 Experimental Results

#### 2009 Clinical Notes Dataset Results

We choose $k = 5$ to cluster documents into 5 groups. For each document clusters, the top 10 features with the highest weight are listed in Table 3.3 (NMF results) and Table 3.4(Multi-view NMF results).

Table 3.3 2009 DATASET RESULTS (NMF)

| # | Symptom | Medication |
|---|---|---|
| 1 | Pain; meds (microcephaly, epilepsy, and diabetes syndrome); infections | Fluvastatin; nicardipine; methyldopa; amphotericin; thera; ammonia; hydroxyzine hcl |
| 2 | Congestive heart failure; coronary artery disease; secondaries (neoplasm metastasis); diabetes | Emtricitabine; potassium citrate; bicalutamide; mcp; dipyridamole |
| 3 | Ischaemia; nausea; congestive heart failure; symptoms | Procaine; hydroxyzine hcl; menthol; dextran 40; linezolid; clopidogrel bisulfate |
| 4 | Hypertension; obesity; asthmatics; pulmonary failure; gout; apnea, sleep apnea syndromes; mental depression; hepatitis b; diabetes mellitus; depressive disorder | - |
| 5 | Erythema; diarrhea; abdominal pain; haematocrit; obesity; wound; place (ocular myopathy with hypogonadism); vomiting | Beta blockers; emtricitabine |

Table 3.4 2009 DATASET RESULTS (MULTI-VIEW NMF)

| # | Symptom | Medication |
|---|---|---|
| 1 | Hyperlipidaemia; hypercholesterolaemia; polycythaemia; gerd; hypertensive disease | Aspirin; Lisinopril; furosemide; phencyclidine; metoprolol |
| 2 | Chest pain; constipation; facial hemiatrophy; pain; food-drug interactions | Heparin, porcine; digoxin; amiodarone; furosemide; warfarin |
| 3 | Place (ocular myopathy with hypogonadism); haematocrit; secondaries (neoplasm metastasis); pain; chest pain | Dextrose; insulin; metoprolol; aspirin; creatinine |
| 4 | Diabetes mellitus; glaucoma; hepatitis c; hepatitis c virus; congestive heart failure | Prednisone; insulin, aspart, human/rdna; acetaminophen; vancomycin; levofloxacin |
| 5 | Diabetes mellitus; depression; diabetes; sleep apnea, obstructive; asthma | Insulin glargine; albuterol; Lisinopril; digoxin; furosemide |

In Table 3.3, all the major features in component 4 are symptom names. While Multi-NMF can get uniform symptom names and medication names for each cluster. The solution provides a way to observe intrinsic patterns between symptom names and medication names in each cluster.

**2014 Clinical Notes Dataset Results**

We choose $k = 3$ and $k = 2$. $k = 3$ represents clustering patients into three groups: the first type is patients who develop Coronary Artery Disease (CAD); the second type is patients who have CAD in their first records; and the third type is patients never develop CAD. The result is shown in Table 3.5.

Table 3.5 2014 DATASET RESULTS ($K = 3$)

| Feature Type | Views | Accuracy(%) | NMI |
|---|---|---|---|
| Count | Words | 40.54 | 0.0228 |
| | Symptom/Medication | 52.03 | 0.1273 |
| | All 3 views | 53.38 | 0.1459 |
| TF-IDF | Words | 35.47 | 0.0020 |
| | Symptom/Medication | 52.36 | 0.1606 |
| | All 3 views | 52.36 | 0.1711 |

$k = 2$ represents clustering patients into two groups: The first type is patients who develop Coronary Artery Disease (CAD) or have CAD in their records, and the second type is patients never develop CAD. The result is shown in Table 3.6.

Table 3.6 2014 DATASET RESULTS ($K = 2$)

| Feature Type | Views | Accuracy(%) | NMI |
|---|---|---|---|
| Count | Words | 57.77 | 0.0198 |
| | Symptom/Medication | 55.07 | 0.0924 |
| | All 3 views | 59.80 | 0.1751 |
| TF-IDF | Words | 53.38 | 0.0034 |
| | Symptom/Medication | 73.31 | 0.1844 |
| | All 3 views | 75.00 | 0.2283 |

In both Table 3.5 and Table 3.6, we use word counts and TF-IDF as features to generate the feature matrices. Using symptom names and medication names have better accuracy and NMI than just using words. Using all 3 views (words, symptom names, and medication names) together can achieve the highest performance.

The results of using all three views are compared between NMF and multi-view NMF are shown in Figure 3.4.

When $k = 3$, using word count as feature shows that multi-NMF achieves about 12% higher accuracy than NMF. It has 14% higher accuracy when using TF-IDF as features. When $k = 2$, using word count as the feature, multi-view NMF has the same accuracy as NMF. While using TF-IDF as features, multi-view NMF has 24% higher accuracy. Multi-view NMF has better performances

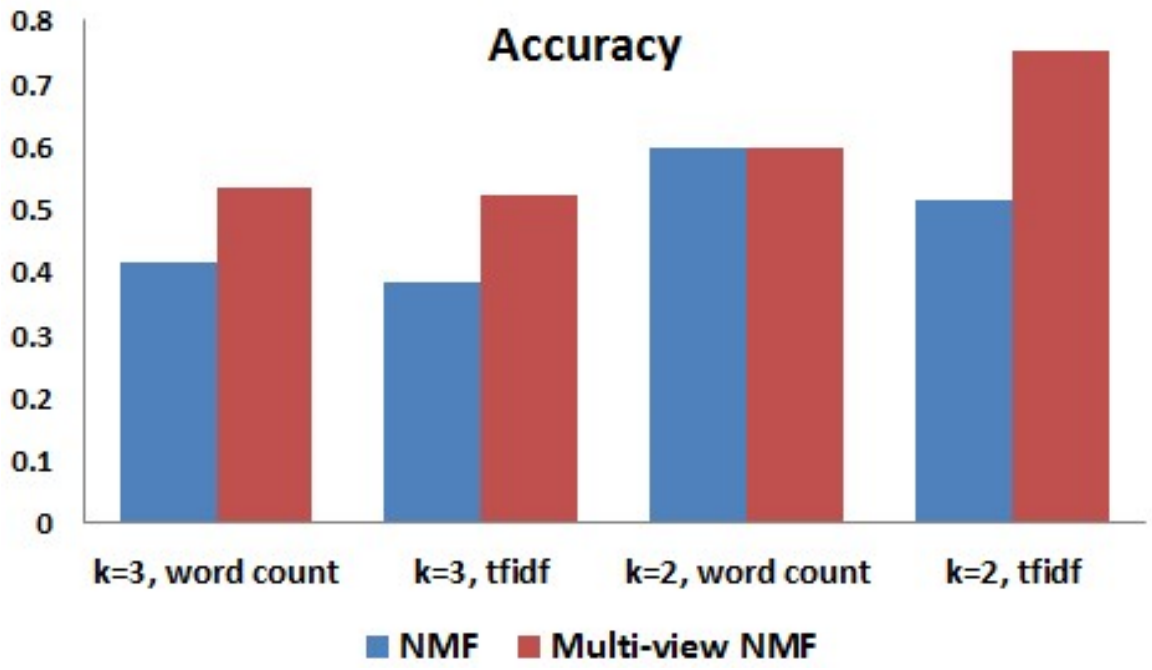



Figure 3.4 Accuracy from Multi-view NMF and NMF.

than NMF.

### 3.5.4 Discussion

In this section, we use extracted symptom/medication names combined with words as three-views from clinical notes and then apply multi-view NMF for documents clustering. Two different datasets are used to compare multi-view NMF with NMF. The 2009 clinical notes dataset presents major features contained in each cluster. For 2014 clinical notes dataset, we use accuracy and NMI as evaluation metrics to compare results. It showed that by using symptom names and medication names, the clustering performance can be improved. It also indicates that multi-view NMF can achieve better results than NMF.

In the future work, we may consider using other information, such as patients age/gender/demographical information, to improve clustering performance; and also explore intrinsic relationships among different views. We also plan to use the document clustering results to improve medication recommendation as discussed in this work [24].

## 3.6 Visualization of Risk Factors for Heart Disease

In this section, we discuss data exploration and visualization of risk factors for heart disease from medical documents using NMF.

### 3.6.1 Motivation

Heart disease can cause severe problems and even death [161], it's a major cause of death among people with diabetes. Patients with diabetes are more likely to have heart disease than patients without diabetes. According to data from Centers for Disease Control and Prevention (CDC) [6], over 20% of people with diabetes aged 35 years or older reporting Coronary Heart Disease (CHD).

Detecting risk factors of heart disease is extremely important in tracking the progression of heart disease in diabetic patients. 2014 i2b2 shared tasks [159] announced the task of identifying risk factors for heart disease and released its datasets of medical records from diabetic patients containing information about heart disease risk factors. This competition [137] achieved a best result of 91.4% precision, 96.8% recall, and 92.8% F-value.

Based on the dataset of annotated medical documents from diabetic patients provided by 2014 i2b2 shared task2 [162], we explore methods to analyze and visualize the risk factors for heart disease in the diabetic population. This study integrates multiple risk factors (hypertension, hyperlipidemia, smoking status, and etc.,) with various attributes into a uniform feature representation, and then applies it to a new framework to cluster and analyze data for patients from multiple aspects. This research work employs NMF to reduce the dimensions and cluster the data for the purpose of visualization. We describe the accuracy of our results and conduct case analysis over results. Our study explores new ways of visually interpreting risk factors for patients and assisting decision making for physicians.

Prior research work about risk factors analysis in the healthcare domain explores the correlations between diseases and some general factors, such as age [163], gender [164], residence information [165], and etc. Some research work further explores the intrinsic relationships among risk factors by using advanced data mining techniques, such as clustering methods and dimension reduction

[6] https://www.cdc.gov/diabetes/statistics/cvd/fig2.htm

methods. For example, clustering methods were used to group the diabetes mellitus population into different clusters, and discover patterns within clusters.

Dimension reduction methods provide solutions for risk factors analysis and visualization. Harle, et al., [166, 166] used dimensionality reduction techniques: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to classify and assess chronic disease risks, and further explored methods for two-dimensional visualization. PCA and LDA provide simple ways to reduce dimensionality.

We explore NMF to provide a more intuitive decomposition of data [158]. NMF is efficient for identifying patterns and discovering classes. In bioinformatics domain, Brunet et al [158] conducted pattern discovery over metagenes and molecular using NMF; Jiang et al [167] applied NMF to research on the functional biogeography of ocean microbes. It has also been applied to document clustering [46], which can capture base topics from document clusters. Prior research about biomedical document clustering [168] showed that background knowledge, such as domain ontology, can improve the quality of document clustering. Zhang, et al., [56] used MeSH ontology to improve the quality of clustering for medical documents.

To the best of our knowledge, no research work has applied NMF methods to analyze risk factors of heart disease for data analysis and visualization from medical documents. Visualization usually provides a better way to present data patterns and amplify recognitions; we explore various ways to visualize data in order to provide some medical insights. NMF methods provide a way to discover underlying patterns among risk factors and cluster patients into "meaningful" classes at the same time. In this section, we explore methods to analyze risk factors extracted from documents, and find solutions to visualize the data. We aim to discuss two research questions:

(1) How to find patterns in data by capturing inherent structures among risk factors and clustering patients into "meaningful" classes?

(2) How to demonstrate and visualize the results generated from the NMF algorithm and to deliver some medical insights?

### 3.6.2 Data Preparation and Analysis

#### Datasets

The overall corpus contains 1304 medical records with risk factors annotated, from 296 patients. Each patient has about 3-5 medical records. These documents are organized in a timeline for each patient. Risk factors are annotated in these medical documents. There are 8 types of risk factors containing 39 underlying attributes in total.

The risk factors and their attributes are summarized in Table 3.7. Each risk factor is processed as an ordinal variable. For example, "Hyperlipidemia" has three attributes: "Mention", "High chol.", and "High LDL". The term, "Mention," is used to represent the indication of a hyperlipidemia/hypercholesterolemia condition from the text in the medical documents. For example, if the text "a diagnosis of Hyperlipidemia" occurs in a medical document, it will be annotated as a "Mention" for the "Hyperlipidemia" variable. As such, a text clips reading "latest LDL: 135" will be annotated as "High LDL" according to the annotation guidelines [162].

We use all the annotated risk factors and their attributes as features to represent patients' conditions. There are 39 different attributes summarized in Table 3.7, so we use a feature vector with 39 dimensions to represent each patient $p_i \in P$. The value of each feature in the vector for patient $p_i$ is the total number of corresponding annotated risk factors which occur in all medical documents for this patient. We use $m$ to represent the total number of patients (in our datasets, $m = 296$). We use n to represent the total number of features ( $n = 39$). We build a matrix $A$ of size $n \times m$ to represent the patient population and their features.

#### An Example

NMF provides a way to decompose and visualize the dataset from a dual view. NMF is applied to factorize matrix $A$ into two matrices $A \sim W \times H$. Matrix $A$ represents the $n$ features in $m$ patients. Matrix $W$ has size $n \times k$, and matrix $H$ has size $k \times m$. $k$ represents a small number of components. The meaning of components can be illustrated from two aspects: (1) for matrix $W$, $k$ components represent underlying patterns among risk factor features; (2) for matrix $H$, each patient has different portions in different components. We can use matrix $H$ to cluster patients into different classes.

Table 3.7 Risk factors and attributes.

| Variables | Attributes | Explanations/Examples |
|---|---|---|
| Diabetes | Mention, A1C, Glucose | Mention: Some phrases indicate patient has diabetes: "has diabetic ketoacidosis"; A1C: A1c test over 6.5;... |
| Coronary Artery Disease (CAD) | Mention, Event, Test, Symptom | Test: test shows ischemia; Symptom: "chest pain consistent with angina"; ... |
| Hyperlipidemia | Mention, High chol., High LDL | High chol.: total cholesterol of over 240; High LDL: LDL over 100mg/dL;... |
| Hypertension | Mention, High bp | Mention: a diagnosis of hypertension; High bp: BP over 140/90mm/hg;... |
| Obese | Mention, BMI | BMI: BMI over 30;... |
| Family_Hist | Present, Not present | Present: first-degree relative was diagnosed as prematurely CAD;... |
| Smoker | Current, Past, Ever, Never, Unknown | Current: "Patient has smoking habit";... |
| Medication | Metformin, Insulin, Sulfonylureas, Thiazolidinedione, DPP4 inhibitors, Anti-diabetes, Aspirin, Thienopyridine, Beta blocker, ACE inhibitor, Ezetimibe, Nitrate, Calcium channel blocker, Statin, Fibrate, Niacin, ARB, Diuretic | Aspirin: "Current medications: Aspirin";... |

Figure 3.5 and Figure 3.6 are generated from Brunet et al.'s [158] source codes of NMF implementation. In Figure 3.5, we set $k = 2$. Matrix $A$ is factorized into $W \times H$. On the top of Figure 3.5, the column of matrix $A$ represents the features' weights for a given patient and the row of $A$ represents the weight of a given feature across patients. There are no obvious patterns in matrix $A$. Colors ranging from dark red, to dark blue, indicated the changing of weight value (Red high, blue low).

On the left of Figure 3.5, the column of matrix $W$ represents portions of features in each component. For example, in component 2, features 4, 5, 6, 7, and 33 have relatively higher weights. These features are CAD-mention, CAD-event, CAD-test, CAD-symptom, and medication-nitrate. All of these features are related to Coronary artery disease (CAD). CAD related features dominate in component 2. Features like diabetes-mention, hyperlipidemia-mention, hypertension-mention, hypertension-high bp, medication-aspirin, and etc. have relatively higher weights in component 1.

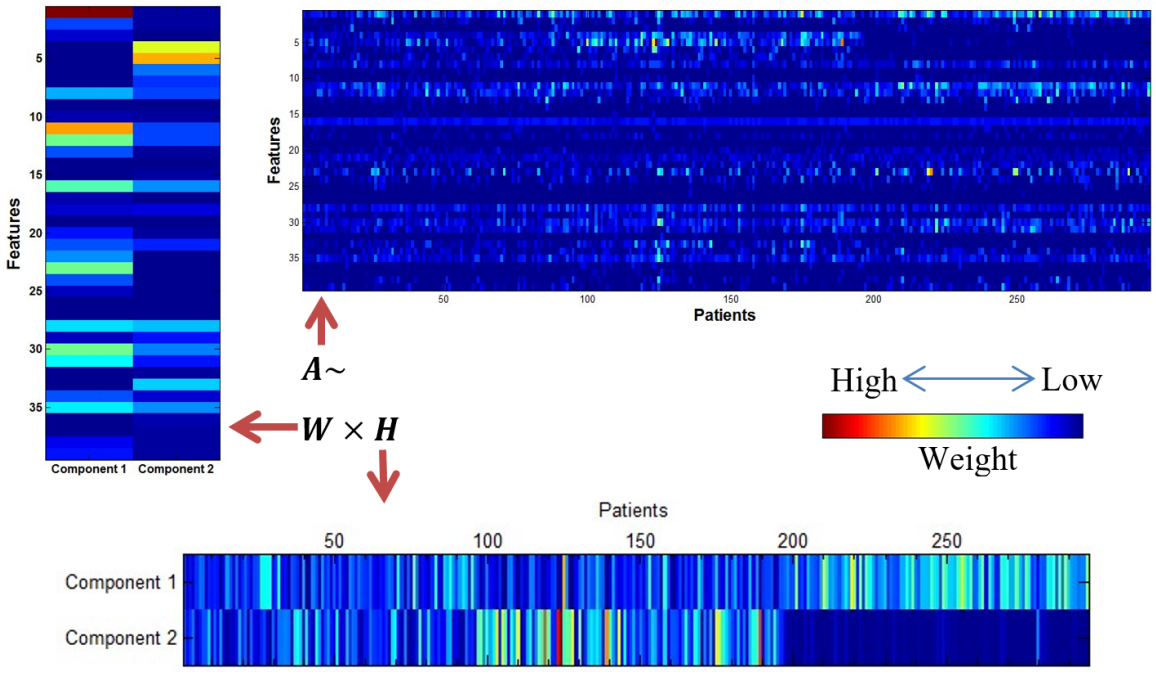



Figure 3.5 An illustration of NMF and results ($k = 2$).

On the bottom of Figure 3.5, the column of matrix $H$ represents the weight of each component at a given patient. The row of $H$ represents the relative weight of a given component across patients.

**Choose $k$ Value**

The NMF algorithm clusters patients into classes and groups feature into components. We use the consensus matrix and cophenetic correlation to decide the $k$ value. The size of the consensus matrix is $m \times m$.

As shown in Figure 3.6, colors ranging from dark red to dark blue indicated clusters' ability for patients to be grouped together (Red high, blue low). These patients are grouped into two clusters clearly at $k = 2$. $k = 4$ also has a relatively clearer clustering result than that of the others. In Figure 3.7, the higher the cophenetic correlation value, the more robust clustering results. Based on the consensus matrices and cophenetic correlation, we choose $k = 2$. Since $k = 4$ also demonstrates relatively robust results in Figure 3.6 and Figure 3.7, we also discuss the results from NMF under $k = 4$.

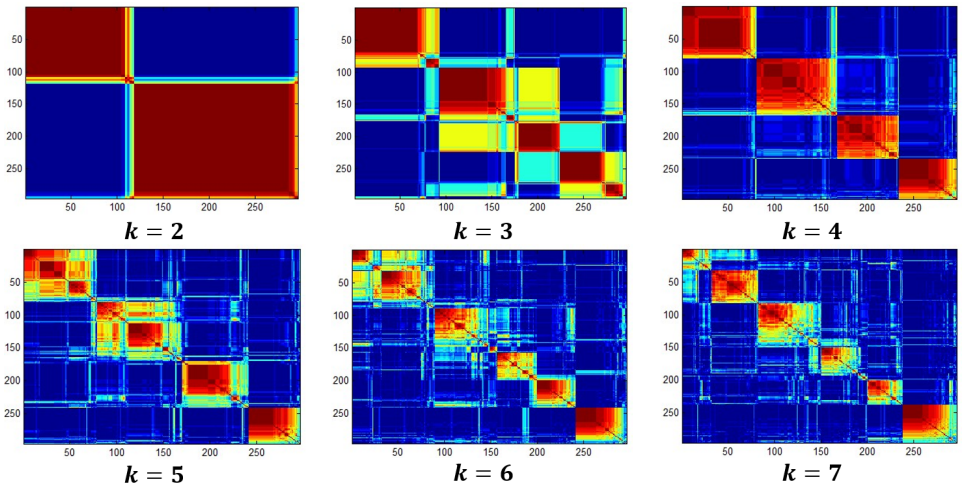



Figure 3.6 Consensus clustering matrices at $k = 2, 3, 4, 5, 6, 7$.

### 3.6.3 Results Analysis and Visualization

**Results Analysis**

When $k = 2$, the initial NMF decomposition results are displayed in Figure 3.5. We use matrix $H$ to classify patients into 2 classes as shown in Figure 3.8.

The patients are clustered into two classes: class 1 and class 2. Class 1 has a higher weight of component 1, and class 2 has a higher weight of component 2. We further visualize the matrix $W$ as shown in Figure 3.9. The horizontal axis represents the 39 features; the vertical axis represents the weight value of each component in each feature; the green bar represents the component 1, and the red bar represents the component 2. The dominating features in component 2 are features: 4, 5, 6, 7, 15, and 33 (CAD-mention, CAD-event, CAD-test, CAD-symptom, FAMILY_HIST-present, and MEDICATION-nitrate). These features have a higher weight in component 2 and lower weight in component 1. The dominating features in class 1 are diabetes-mention, hyperlipidemia-mention, hypertension-mention, hypertension-high bp, medication-aspirin, and etc. Since all these features are more related to CAD than other features and class 2 has a higher weight of component 2. We classify the class 2 as a "high risk" patient class for heart disease. Class 1 is a "low risk" patient class for heart disease.

The original medical documents indicate two types of patients. The first type is patients never

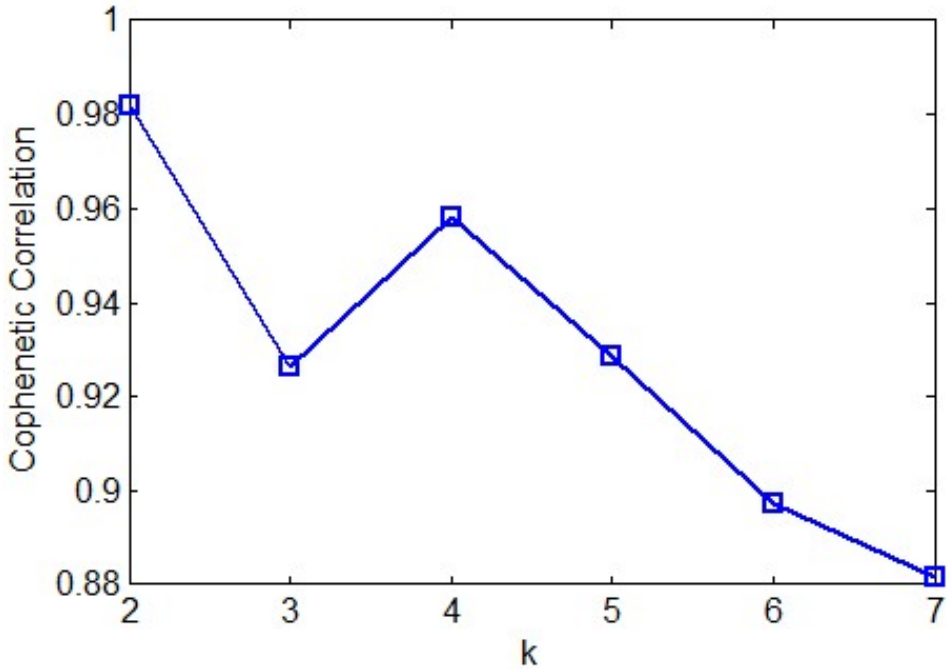



Figure 3.7 Cophenetic correlation result at $k = 2, 3, 4, 5, 6, 7$.

develop CAD; the second type is patients who develop CAD or have CAD in their medical records. We use this as the gold standard to evaluate the accuracy of our results as shown in Table 3.8. 87.8% of Type 2 patients have been grouped into class 2 as a "high risk" class, and 99.0% of Type 1 patients have been grouped into class 1 as a "low risk" class.

Table 3.8 Accuracy of our results.

| | Type1 | Type2 |
|---|---|---|
| Accuracy | 99.0% | 87.8% |

**Case Analysis**

We randomly pick two patients as samples from two classes. These two patients are highlighted in Figure 3.8. The risk factors from these two patients' medical documents are summarized in Table 3.9. Since the risk factors of CAD frequently occur in patient 2's medical documents, it shows that patient 2 probably has a higher chance to have heart disease than patient 1. In Figure 3.8, patient 1 is classified in class 1 (i.e. "low risk" class), and patient 2 (i.e. "high risk" class) is classified

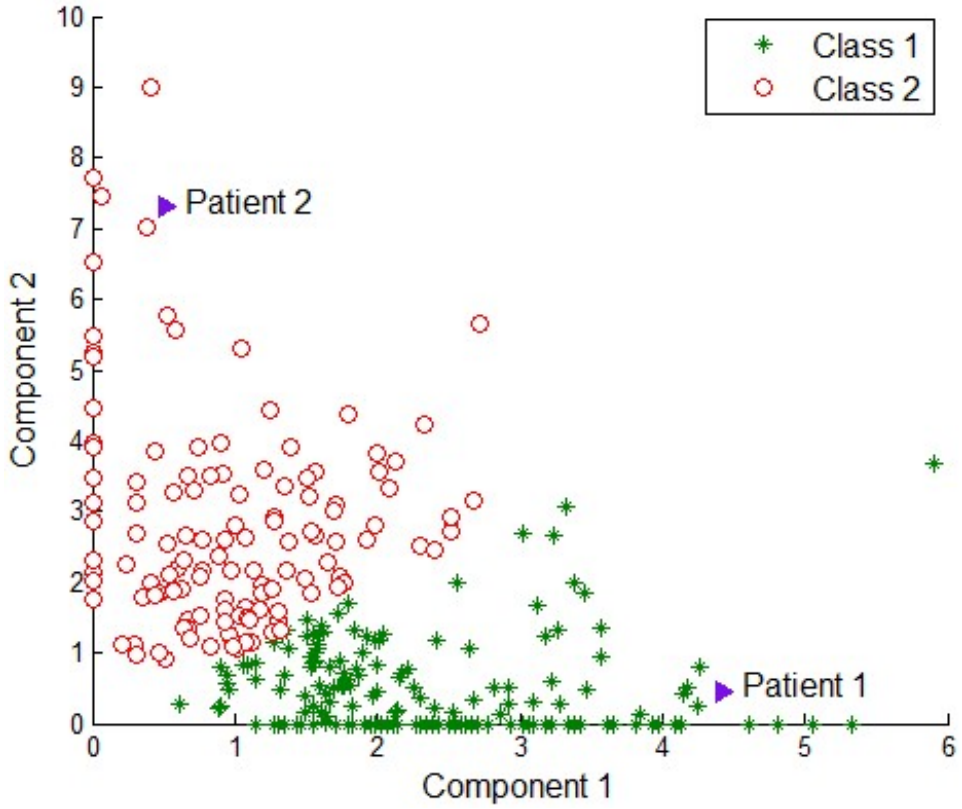



Figure 3.8 Patients Clustering Result at $k = 2$.

in class 2.

Since $k = 4$ also indicates a robust cluster results in Figure 3.6 and Figure 3.7, we use NMF result to cluster patients into four classes. The dominating features for each class of patients are summarized in Table 3.10. Class A has dominating features highly related to CAD, class B has dominating features relatively weaker related to CAD than class A, class C has dominating features related to diabetes, and class D has dominating features related to hyperlipidemia and hypertension. When we pick $k = 2$, patients can be grouped into two classes: class 1 ("low risk" class) and class 2 ("high risk" class). Class A and class B can be regarded as sub-classes of class 2; class C and class D are sub-classes of class 1.

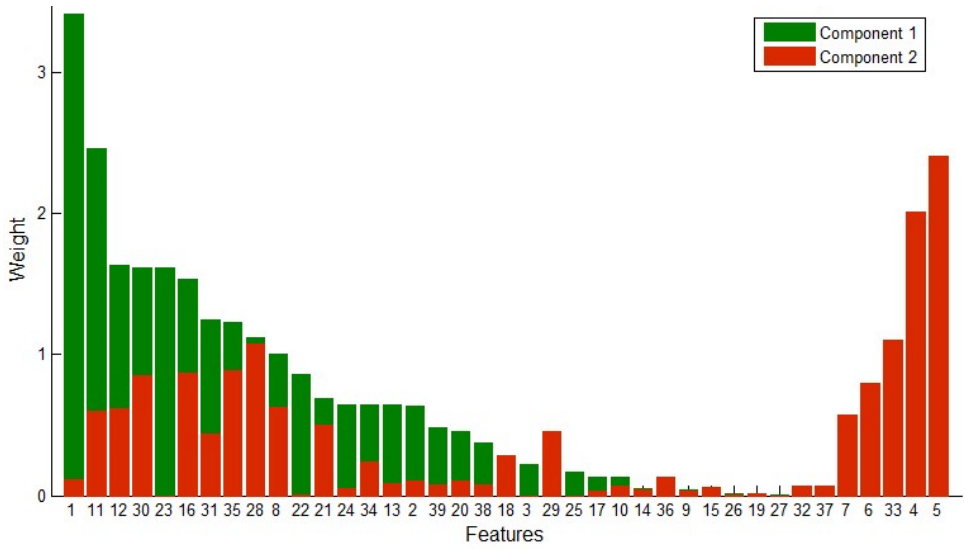



Figure 3.9 Feature Analysis for Patients Clustering Result at $k = 2$.

### 3.6.4 Discussion

In this section, we use NMF-based methods to analyze and visualize risk factors of heart disease for the diabetic population from medical documents. NMF methods provide a way to discover underlying patterns among risk factors and cluster patients into "meaningful" classes at the same time. We build a features-patients matrix to represent the population of patients and risk factors of heart disease, and then describe the process of model selection and matrix decomposition using the NMF algorithm. Based on the NMF results, we discuss the different patient classes and their different features of risk factors. We provide insights on how to cluster patients into classes by capturing inherent patterns among risk factors; and exploration on how to visualize the risk factors of heart disease for the diabetic population.

We only explore the risk factors analysis and visualization at the population level. It is also interesting to explore the visualization of risk factors for each patient from a longitudinal view. We plan to explore methods to integrate timeline information for data analysis and visualization. Except for annotated risk factors, we also plan to extract more additional features from the raw text of medical documents to improve our results.

Table 3.9 The frequency of risk factor being annotated in medical documents for two patients.

| Patient 1 | Patient 2 |
|---|---|
| • Hypertension-mention: 19;<br>• medication-ace inhibitor: 18;<br>• medication-beta blocker: 17;<br>• hypertension-high bp: 12;<br>• medication-aspirin: 11;<br>• obese-mention: 8;<br>• medication-statin: 5;<br>• hyperlipidemia-mention: 5;<br>• family_hist-not present: 5;<br>• medication-diuretic: 4;<br>• smoker-past: 3;<br>• ... | • CAD-event: 19;<br>• CAD -mention: 17;<br>• hypertension-mention: 10;<br>• hypertension-high bp: 8;<br>• medication-statin: 8;<br>• hyperlipidemia-mention: 7;<br>• medication-diuretic: 7;<br>• medication-ACE inhibitor: 6;<br>• medication-aspirin: 6;<br>• medication-nitrate: 6;<br>• medication-nitrate: 5;<br>• family_hist-not present: 5;<br>• smoker-past: 5;<br>• medication-calcium channel blocker: 5;<br>• hyperlipidemia-high LDL: 3;<br>• obese-BMI: 3;<br>• medication-thienopyridine: 3;<br>• ... |

## 3.7 Conclusion

In this chapter, we present a system for extracting symptom/medication names from clinical notes. Based on extracted concepts, we apply NMF and multi-view NMF to evaluate the effects of using medication/symptom names to improve the clinical documents clustering results. We use NMF to cluster clinical notes into meaningful clusters based on sample-feature matrices. Our experimental results show that multi-view NMF is a preferable method for clinical document clustering. Moreover, we find that using extracted medication/symptom names to cluster clinical documents outperforms

Table 3.10 The dominating features in each patient class (k=4).

| Patients | Dominating Features |
|---|---|
| Class A | CAD-mention; CAD-event; CAD-test; medication-aspirin; medication-thienopyridine |
| Class B | CAD-symptom; smoker-unknown; medication-nitrate |
| Class C | diabetes-mention; diabetes-A1C; diabetes-glucose; medication-insulin |
| Class D | hyperlipidemia-mention; hypertension-mention; hypertension-high bp; family_hist-not present; medication-metformin; medication-sulfonylureas; medication-beta blocker; medication-ACE inhibitor; medication-calcium channel blocker; medication-statin; medication-ARB; medication-diuretic; etc. |

just using words. We also explore to use NMF for data exploration and visualization of risk factors for heart disease from medical documents.

# Chapter 4: Symptom/Medication Relation

In this chapter, we present our work on symptom/medication relation extraction from clinical documents. Symptom and medication information existing in clinical notes are valuable. Little research has been done on matching medication information with multiple symptoms information. Such a matching could provide valuable information for patients with multiple syndromes. We propose a Symptom-Medication (Symp-Med) matching framework to model symptom and medication relationships from clinical notes. After extracting symptom and medication concepts, we construct a weighted bipartite graph to represent the relationships between the two groups of concepts. The key is to efficiently answer user's symptom-medication queries using the graph. We formulate this problem as an Integer Linear Programming (ILP) problem. The objectives are to maximize the total edge weight and minimize the number of medication concepts. We first explore a Branch-and-Cut based algorithm. Then, we revise the combinational objective and propose a Greedy-based algorithm for solving the Symp-Med problem. The Greedy-based algorithm performs better and significantly improves the computational costs.

## 4.1 Motivation

Symptoms and medications are two important types of information that can be obtained from clinical notes. Symptom information such as *diseases*, *syndromes*, *signs*, *diagnose* etc., can be used to analyze diseases for patients. In addition, valuable medication information is commonly embedded in unstructured text narratives spanning multiple sections in medical documents [153]. Medication information from clinical notes is often expressed with medication names and other signature information about drug administration, such as *dosage*, *route*, *frequency*, and *duration*. In this section, we extract medication names from clinical notes, and use medication names as medication concepts. Other related medication information is also very important, and will be considered in future research.

Currently, large volumes of clinical documents are generated by electronic health record systems [154]. On one hand, these clinical documents are unstructured or semi-structured. It is a difficult task to extract information from these documents. Symptom information and medication information extraction for clinical notes need sophisticated clinical language processing methods [10]. On the other hand, due to the individual diversity, discovering and mining relationship between symptom information and medication information from clinical texts becomes a challenging problem. These underutilized resources have a huge potential to improve health care. It is very important for patients with multiple syndromes to learn the relationships between symptoms and medications as indicated in the scenario below.

A use case scenario: a new patient is diagnosed with an alcoholic liver disease (ALD) and type2 diabetes. A set of related symptoms are observed, so a set of medications should be prescribed to treat these symptoms. In the meantime, related clinical notes extracted from a database with symptoms and medications highlighted will also be presented as evidence to the physician and patient. The physician can use these clinical notes to support decisions, and the patient might find the medications given by physician more convincing based on the clinical notes from other patients who had similar medical conditions.

In this section, we study the following questions:

- How to represent the relationship of symptom concepts and medication concepts we extracted from clinical notes?
- How to extract a set of most valuable medication concepts for a patient with a set of known symptom concepts?

To the best of our knowledge, little previous work has systematically studied these problems.

## 4.2 Symp-Med Framework

Base on the symptom concepts and medication concepts extracted from clinical notes, we develop a Symp-Med Framework. The major component of this framework is a Symptom and Medication Bipartite Graph (Symp-Med Bi-graph).

### 4.2.1 Symp-Med Graph

The Symp-Med Bi-graph is a bipartite graph $G = \langle S \cup D, E \rangle$. There are two groups of nodes $S$ and $D$. There is no edge between vertices in the same group. $S$ is a set of vertices representing symptom concepts from clinical notes, $S = \{s_i | 1 \leq i \leq p\}$. $D$ is a set of vertices representing medication concepts from clinical notes, $D = \{d_i | 1 \leq i \leq q\}$. $E$ is a set of edges between the vertices from $D$ and $S$, $E \in S \times D$. $M$ is a set of weights representing weight value for each edge in set $E$.

The Symp-Med Bi-graph $G$ can be represented by a $p \times q$ dimension matrix $M$, where $m_{ij}$ is the weight value of edge $\langle s_i, d_j \rangle$. For each clinical note, we use the symptom and medication concepts to form a matrix $M$. We set the value of $m_{ij}$based on the relation information we extracted from the clinical note. We aggregate all matrix $M$ for individual clinical notes (in the clinical note level) to form a new matrix $W$ for all clinical notes (in the cluster level).

### 4.2.2 Weight Matrix Definition

For a clinical note, we extract a set of symptom concepts $S = \{s_i | 1 \leq i \leq p\}$ and a set of medication concepts $D = \{d_j | 1 \leq j \leq q\}$. A matrix $M_{p \times q}$ can be built based on these two sets of concepts. We define a weight factors set $F = \{f^r | 1 \leq r \leq k\}$, which contains multiple weight factors. The weight factor set decides the weight values for each concepts pair $\langle s_i, d_j \rangle$. Weight values represent the relevance between symptom concept and medication concept. The larger the weight values, the more relevant the two concepts are. The weight value $m_{ij}$ for concept pair $\langle s_i, d_j \rangle$ with weight factor value is defined as follows:

$$m_{ij} = \sum_{k}^{r=1} f_r^{ij} \tag{4.1}$$

We define two types of weight factors. One is a "Co-occurrence" factor $f_{ij}^1$. If symptom concept $s_i$ and medication concept $d_j$ appear in the same clinical note, $f_{ij}^1 = 1$. Otherwise, $f_{ij}^1 = 0$. The second weight factor is a "Co-occurrence in the same section" factor $f_{ij}^2$. If symptom concept $s_i$ and medication concept $d_j$ appear in the same section of a clinical note, $f_{ij}^2 = 1$. Otherwise, $f_{ij}^2 = 0$.

For all clinical notes $C = \{c_i | 1 \leq i \leq k\}$, a matrix $W$ for all clinical notes $C$ is constructed by

integrating all weight matrices $M$.

## 4.3 Symp-Med Matching Algorithm

In the weight matrix $W$ learned from the Symp-Med framework, the weight values represent the relevance relations between symptom concepts and medication concepts. For the Symp-Med framework, we define the Symp-Med matching problem. For a set of symptom concepts from a patient as the input, we want to predict a set of medication concepts as the output with the maximized total edge weight value and minimized the number of medications. A motivating example for our Symp-Med matching problem is illustrated as follows.

A patient has two symptoms: *fever* and *runny nose*. A physician may have two kinds of prescriptions for this patient. The first prescription contains one medication, "*Compound Paracetamol and Amantadine Hydrochloride Tablets*". The second prescription contains two medications, "*Acetaminophen*" and "*Nasal Drops*". Suppose the first prescription has a higher weight value with these two symptoms than the second prescription. First, set 1 (*Compound Paracetamol and Amantadine Hydrochloride Tablets*), and set 2 (*Acetaminophen* and *Nasal Drops*) should be matched as two medication sets for these two symptoms. Second, since the first prescription "*Compound Paracetamol and Amantadine Hydrochloride Tablets*" has the larger weight value and a smaller number of medications, it should be matched as the top one in the output set.

### 4.3.1 Symp-Med Matching Problem Formulation

We formulate the Symp-Med matching problem as follows.

**Input**

For this Symp-Med matching problem, the input includes a weight matrix $W$ and a query vector $S^{'}$. The weight matrix $W$ is a $m \times n$ dimension matrix. The matrix describes the weight values of relevance edges between a set of symptom concepts $S = \{s_1, \ldots, s_m\}$ and a set of medication concepts $D = \{d_1, \ldots, d_n\}$. The query vector $S^{'}$ is described as follows:

$$S^{'} = \{s_1^{'}, \ldots, s_p^{'}\}, \tag{4.2}$$

$p \leq m$, $S^{'} \subseteq S$, where $i, j \in \{1, 2, \ldots, p\}$, and $\forall i \neq j$, $s_i^{'} \neq s_j^{'}$

**Output**

Given the weight matrix $W$ and query vector $S^{'}$, we want to get a set of medication concepts as output, which can be represented as a vector as follows:

$$D^{'} = \{d_1^{'}, \ldots, d_q^{'}\}, \tag{4.3}$$

$q \leq n$,$D^{'} \subseteq D$, where $i, j \in \{1, 2, \ldots, q\}$, and $\forall i \neq j$,$d_i^{'} \neq d_j^{'}$

**Constraints and Goal**

The solution is a sub matrix of $W$ for the query vector $S^{'}$ and the output vector $D^{'}$. This sub matrix $W^{'}$ is $p \times q$ dimension matrix. In order to guarantee that the summation value of all elements from one row in matrix $W^{'}$ is bigger than zero, a constraint is set as follows:

$$\sum_{j \in \{1,\ldots,q\}} w_{ij}^{'} > 0, \tag{4.4}$$

$$\text{for any } i \in \{1, \ldots, p\}$$

That means there is at least one element larger than zero in each row since all weight values are either equal to zero or larger than zero.

The goal of this problem is two-fold:

First, maximize the sum of all elements (total weight value) from Matrix $W^{'}$, which is described as follows:

$$\sum_{i \in \{1,\ldots,p\}, j \in \{1,\ldots,q\}} w_{ij}^{'} \tag{4.5}$$

Second, minimize the number of columns $q$. That means the size of output vector should be as small as possible.

### 4.3.2 Symp-Med Matching Algorithm

First, the Symp-Med matching problem can be formulated as an ILP problem, the form of this ILP problem is described as follows:

$$\max \sum_{i=1}^{p} \sum_{j=1}^{n} w_{ij} z_{ij} - \epsilon \sum_{j=1}^{n} y_j \tag{4.6}$$

$$\textit{Subject to}$$

$$\sum_{j=1}^{n} z_{ij} \geq 1, \forall i \in \{1, \ldots, p\}$$

$$z_{ij} \leq y_j, \forall i \in \{1, \ldots, p\}, \; z_{ij} \leq y_j, \forall j \in \{1, \ldots, n\}$$

$$z_{ij} \in \{0, 1\}^{p \times n}$$

$$y_j \in \{0, 1\}^{n}$$

Equation 4.6 uses $z_{ij}$ and $y_j$ to decide whether an element $d_j \in D$ should be selected to $D^{'}$ or not. $z_{ij} = 1$ means that the edge $\langle s_i, d_j \rangle$ is selected. $z_{ij} \leq y_j, \forall i \in \{1, \ldots, p\}, \forall j \in \{1, \ldots, n\}$ means if any edge connect with $d_j$ is selected, $d_j$ need to be selected. $y_j = 1$ represents $d_j$ is selected. If none of edges connecting to $d_j$ is selected, $d_j$ is not selected, then $y_j = 0$.

$\epsilon$ is a parameter to balance the two objectives: $\sum_{i=1}^{p} \sum_{j=1}^{n} w_{ij} z_{ij}$ and $\sum_{j=1}^{n} y_j$. $\epsilon$ is set dynamically as follows:

$$\epsilon = \epsilon^{'} \times \max_{i} (\sum_{j=1}^{n} w_{ij}), \tag{4.7}$$

$$i \in \{1, \ldots, p\}, \; \epsilon^{'} \in (0, 1]$$

$\epsilon^{'} = 1$ represents minimizing $\sum_{j=1}^{n} y_j$ as much as possible, if constraints are all satisfied, no extra $d_j$ will be selected. If $\epsilon^{'} > 1$, the result is the same as the result when $\epsilon^{'} = 1$. The decrease of $\epsilon^{'}$ from one to zero will improve the number of selected $d_j$. When $\epsilon^{'} = 0$, the minimizing objective $\sum_{j=1}^{n} y_j$ is not considered. In order to take the maximized total weight value and the minimized selected $d_j$ number both into consideration, $\epsilon^{'}$ is set as $\epsilon^{'} \in (0, 1]$.

The ILP problem formulated in Equation 4.6, which is an NP-hard problem. Approximation algorithms are developed for dealing with ILP problem, such as Primal-Dual method [169], and

Linear Programming (LP) relaxation and rounding method. Here we use a branch-and-cut algorithm [170] to solve the ILP problem. The branch-and-cut algorithm is implemented in GLPK MIP solver [171].

Algorithm 1 Branch-and-Cut based Symp-Med Matching

1: **Input:** weight matrix $W \in R^{p \times n}$, parameter $\epsilon^{'}$
2: **Output:** vector $D^{'}$
3: Let $ILP_{SMM}$ be the linear integer programming formulation as Equation 4.6
4: $Y \leftarrow branch\_and\_cut(ILP_{SMM})$
5: **for** $y_j \in Y$ **do**
6: **if** $y_j = 1$ **then**
7: $D^{'} = D^{'} \bigcup \{d_j\}$
8: **end if**
9: **end for**

The branch-and-cut algorithm needs to relax the $ILP_{SMM}$ to a corresponding $LP_{SMM}$. The computational effort to solve LP is bounded by a polynomial function of problem size. The problem size of this $LP_{SMM}$ is $(p+1)n$. A possible computational complexity is $O(pn^2)$ [172].

Since the two objectives in the Symp-Med matching problem are maximizing $\sum_{i=1}^{p}\sum_{j=1}^{n} w_{ij}z_{ij}$ and minimizing $\sum_{j=1}^{n} y_j$ at the same time, then the objective can also be represented as:

$$Maximize \frac{(\sum_{i=1}^{p}\sum_{j=1}^{n} w_{ij}z_{ij})}{(\sum_{j=1}^{n} y_j)} \tag{4.8}$$

The objective in Equation 4.8 is maximizing the unit weight values for each selected $d_j$ in $D^{'}$. Equation 4.8 has the same constraints in Equation 4.6. Since the final output of the Symp-Med matching problem is a vector $D^{'}$ with maximized unit weight value. An optimal result can be obtained in polynomial time without solving $z_{ij}$ and $y_j$ in Equation 4.8. A Symp-Med Matching algorithm based on a greedy method is designed to solve this problem.

Alg. 2 applies greedy method. It uses a score vector $A$ to sort $d_j \in D$ in descending order, and an index vector $B$ to indicate if the constraint Equation 4.4 is satisfied or not. It incrementally extends $D^{'}$ until all the constraints are satisfied.

Algorithm 2 Greedy-based Symp-Med Matching

**Input:** weight matrix $W \in R^{p \times n}$, parameter $\epsilon^{'}$
**Output:** vector $D^{'}$
score vector $A \in R^{1 \times n}$
index vector $H \in R^{1 \times n}$ stores indexes for elements in $D$ sorted in descending order according to $A$
index vector $B \in R^{1 \times p}$
**for** $b_i \in B$ **do**
    $b_i = false$
**end for**
**for** $a_j \in A$ **do**
    $a_j = \sum_{i=1}^{P} w_{ij}$
**end for**
$H \leftarrow sort(A)$
**for** $h_j \in H$ **do**
    **for** $b_i \in B$ **do**
        **if** $b_i = false$ and $w_{ij} > 0$ **then**
            $D^{'} = D^{'} \bigcup \{d_{h_j}\}$
            $b_i = true$
        **end if**
    **end for**
**end for**

## 4.4 Experiments

The motivation of our experiments is two-fold: (1) To examine how the value $\epsilon^{'}$ affect the performance of Branch-and-Cut based Symp-Med Matching Algorithm; (2) To evaluate the performance of Greedy-based Symp-Med Matching algorithm. The rest of this section presents a detailed description of our dataset, experimental design, evaluation methodology, and result analysis.

### 4.4.1 Dataset Description and Evaluation Methodology

We use the clinical notes dataset from the 2009 i2b2 workshop on NLP challenges [17] as experiment dataset. There are 1249 clinical notes in total. After pre-processing, 1239 clinical notes remain. We divided the dataset into 4 groups randomly. Each group has a training set and test set. In each group, 155 clinical notes are used as the training set, and 155 clinical notes are used as the test set in each group. We extract about 1215-1346 symptom concepts and 609-664 medication concepts for each training/test set.

We evaluate the accuracy of algorithms using two sets of evaluation metrics: 1) Precision (P)

and Recall (R); 2) True Positive Rate (TPR) and False Positive Rate (FPR) [173]. ROC (receiver operating characteristic) curve shows how the true positive rate varies with the false positive rate. The area under the ROC curve (AUC) presents achievable TPR with respect to FPR.

### 4.4.2 Symp-Med Matching Analysis

By varying the value of $\epsilon^{'}$, we obtain average performance results of Branch-and-Cut based Symp-Med Matching from four groups of datasets. The result is shown in Table 4.1.

Table 4.1 AVERAGE PERFORMANCE RESULTS OF Alg. 1

| $\epsilon^{'}$ | TPR | FPR | Precision | Recall |
|---|---|---|---|---|
| 0.1 | 0.558 | 0.080 | 0.208 | 0.558 |
| 0.2 | 0.397 | 0.032 | 0.311 | 0.397 |
| 0.3 | 0.313 | 0.018 | 0.396 | 0.313 |
| 0.4 | 0.225 | 0.009 | 0.494 | 0.225 |
| 0.5 | 0.162 | 0.005 | 0.553 | 0.162 |
| 0.6 | 0.133 | 0.003 | 0.587 | 0.133 |
| 0.7 | 0.110 | 0.003 | 0.589 | 0.110 |
| 0.8 | 0.089 | 0.002 | 0.582 | 0.089 |
| 0.9 | 0.068 | 0.002 | 0.614 | 0.068 |
| 1.0 | 0.048 | 0.001 | 0.634 | 0.048 |

$\epsilon^{'}$ is used to balance the objective of maximizing the total weight value and minimizing the total selected $d_j$ number. $\epsilon^{'} = 1$ means only adding necessary $d_j$ to result sets, because each time adding a new $d_j$, it costs the value of $\max_i(\sum_{j=1}^{n} w_{ij})$ loss to the total maximum objective function. So when $\epsilon^{'} = 1$, it achieves the largest precision, but the smallest recall. The average experiment precision is 63.4%, and recall is 4.8%. By decreasing the $\epsilon^{'}$ value, the precision decreases, but the recall increases. When $\epsilon^{'} = 0.1$, we have the lowest precision, 20.8%, and highest recall, 55.8%. When $\epsilon^{'} = 0$, the objective of minimizing selected $d_j$ number is not considered. The algorithm returns all the $d_j$ in $D$ which connects to any element in $S^{'}$. In our experiments, the average precision is 3.74%, and the average recall is 99.7%.

We implement Greedy-based Symp-Med Matching on the four groups of datasets, and the average results are in Table 4.2.

The objective of Alg. 2 is to maximize the unit weight values. The average precision is 63.4%, and the average recall is 6.1%. The result is close to the result in Table 4.1 when $\epsilon^{'} = 1$. The Alg. 1

Table 4.2 AVERAGE PERFORMANCE RESULTS OF Alg. 2

| TPR | FPR | P | R |
|---|---|---|---|
| 0.061 | 0.001 | 0.634 | 0.061 |

can capture the full spectrum of performances by varying the value of $\epsilon^{'}$, while Alg. 2 can produce a good precision result and improve the recall without solving the corresponding LP problem in Alg. 1.

We only remove negation concepts by negation annotator and section annotator during pre-processing. There are a lot of noises exist in extracted symptom and medication concepts. Based on the most frequent sections with symptom and medication concepts, we implement our algorithms on the datasets only contain symptom concepts from most frequent sections in clinical notes. Let indicate the experiments on selected sections from clinical notes as Set 2 experiment, and the experiments on all sections as Set 1. The results in Table 4.1 and Table 4.2 are from Set 1 experiment.

We use ROC curves and Precision-Recall (PR) curve to capture the full spectrum of performances of Set 1 and Set 2 experiments as shown in Figure 4.1.

As shown in Figure 4.1(a), the ROC curve indicates the set 2 has a better result than set 1, since the AUC is slightly larger in set 2. Both set 2 and set 1 have better performance than the Random Guess result. In Figure 4.1(b), the result indicates set 2 also has a better precision/recall results than set 1. The performances of Symp-Med matching algorithms can be improved if more noises can be removed from extracted symptom and medication concepts in the pre-processing stage.

## 4.5 Conclusion

In this chapter, we present a Symp-Med matching framework for representing and mining relationships between the symptom and medication extracted from clinical notes. We formulate the Symp-Med matching problem as an ILP problem and propose Symp-Med matching algorithms for solving the Symp-Med matching problem. We explore a Branch-and-Cut based Symp-Med matching algorithm to solve the ILP problem and define a parameter to balance the two objectives in the ILP

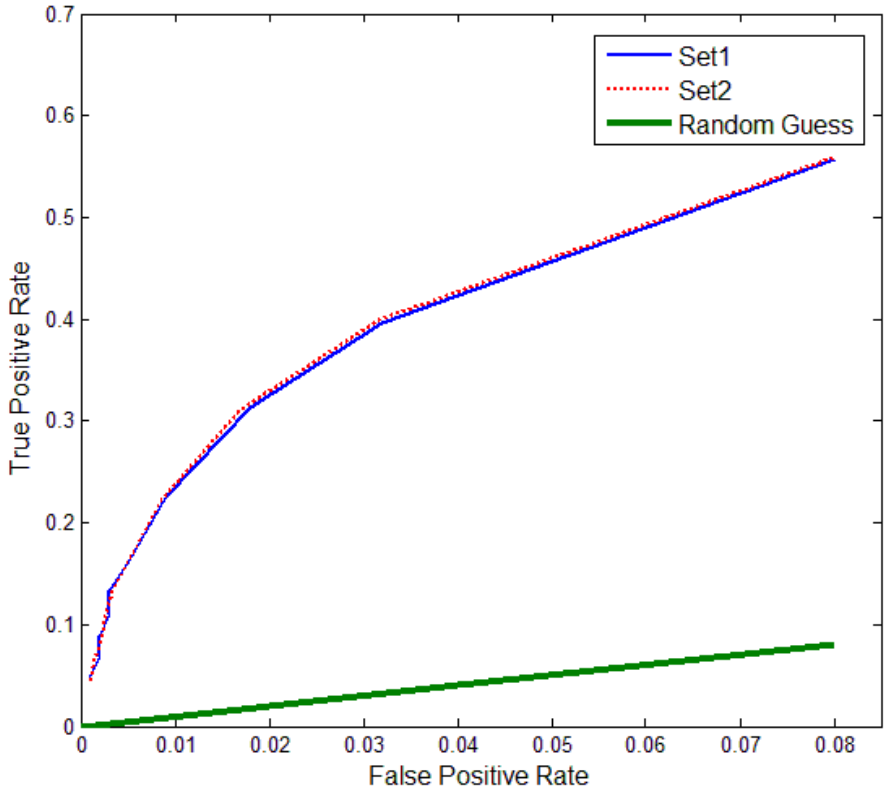
problem. Then we change the objective function in the ILP problem to a combined maximizing the unit weight value objective and propose a Greedy-based Symp-Med matching algorithm for solving it.

Our Symp-Med matching algorithms can be used to predict a set of medications based on a given symptom set. The Symp-Med matching framework can also be applied to error detection [4] for medications in clinical notes. In future work, we plan to improve current work from the following aspects:

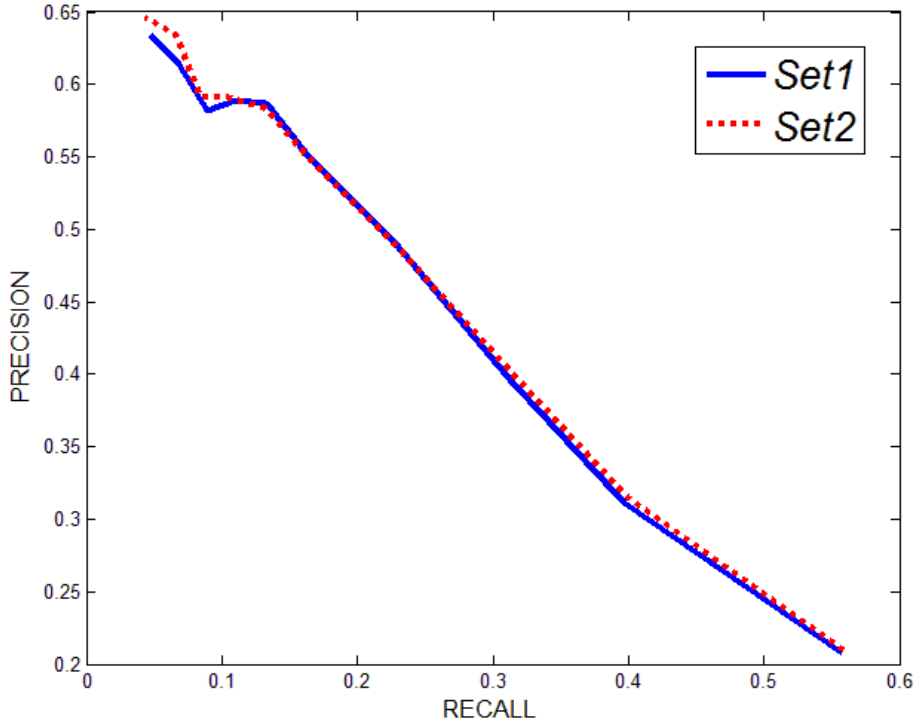
1) We build a Symp-Med weight matrix for our Symp-Med framework. We intend to extend to the weight factor set. Currently, we only use the information extracted from experiment clinical notes dataset to build the weight factor set. Only two weight factors are defined in this paper. In the future, we plan to integrate other factors into the weight factor set, such as drug indications, side effects of drugs, drug interactions, drug administration information etc., from publicly available datasets such as DrugBank, RxNorm, and UMLS etc. [174]

2) There are still a lot of noises remained in extracted symptom concepts and medication concepts during clinical notes pre-processing. These noises affect the performance of our Symp-Med matching algorithms. Improving the results of symptom and medication extraction is worthwhile.

3) Currently, we only consider the relationship between symptom concepts and medication concepts. We plan to integrate symptom-symptom and medication-medication relationships into the Symp-Med framework. For example, we plan to use similarity to build a symptom-symptom matrix. This will help to expand and discover more related symptom information for patients based on observed symptoms.




(a) ROC curve




(b) PR curve

Figure 4.1 Comparison in ROC and PR Curves.

# Chapter 5: Word Embedding Models for Clinical NLP

Word embedding in the NLP area has attracted increasing attention in recent years. The continuous bag-of-words model (CBOW) and the continuous Skip-gram model (Skip-gram) have been developed to learn distributed representations of words from a large amount of unlabeled text data (as discussed in section 2.4.2). In this chapter, we explore the idea of integrating extra knowledge to the CBOW and Skip-gram models and applying the new models to biomedical NLP tasks [175]. The main idea is to construct a weighted graph from knowledge bases (KBs) to represent structured relationships among words/concepts. In particular, we propose a GCBOW model and a GSkip-gram model respectively by integrating such a graph into the original CBOW model and Skip-gram model via graph regularization. Our experimental results on both standard datasets and biomedical NLP tasks show encouraging improvements with the new models. Moreover, the evaluations on two biomedical NLP tasks, biomedical similarity/relatedness task and biomedical information retrieval (IR) task, show that our methods have better performance than baselines.

## 5.1 Introduction

Distributed word representations for solving NLP problems have attracted much attention [78, 79, 80, 81, 82, 83, 84]. In contrast to traditional one-hot representation, which has the limitation of representing implied semantic relations among words, distributed representation uses a dense and low dimensional vector to represent a word. Similar words will be transferred into similar vector representations. It can capture semantic information among words. Mikolov et al. [85, 86] proposed two embedding methods: the continuous bag-of-words model (CBOW) and the continuous Skip-gram model (Skip-gram). They have attracted a great deal of attention among NLP researchers and practitioners [87, 88, 89].

However, the embedding models still have some limitations. First, training a good word embedding model generally requires a very large text corpus. Second, the unlabelled text corpus may

contain noises for learning. For example, words may have incomplete and ambiguous meanings. Recently, some researchers have attempted to encode extra knowledge into word embedding models [91, 92, 93, 94, 95, 96, 90, 97]. One frequently mentioned knowledge resource for enhancing word embedding models is structured Knowledge Base (KB). We have witnessed a quick development of KBs in past years. KBs store structured information about entity types and relation triples. A triple is represented as $\langle subject, predicate, object \rangle$. Many large-scale KBs of general or specific domains have been constructed, such as WordNet [100], Yago [123], Freebase[101], DBpedia [124], and NELL [125], UMLS [29]. KBs are useful resources and powerful tools for supporting NLP tasks such as relation extraction [126, 127] and question answering [118].

In the biomedical domain, there is a growing number of studies on applying word embedding models to biomedical NLP tasks. Tang et al [176] studied the effect of word embedding features on biomedical named entity recognition tasks. Muneeb et al [75] evaluated word similarity task using two word embedding models: word2vec and GloVe. The effect of input corpus and all kinds of parameters for word embedding models are systematically evaluated on biomedical NLP tasks [177, 178, 179]. The parameters include negative sample size, learning rate, vector dimension, context window size, and etc. In spite of the fact that KBs play an important role in biomedical NLP tasks [34, 16], to the best of our knowledge, there is little work on integrating KBs with word embedding models for biomedical NLP tasks.

In this chapter, we explore the idea of using extra knowledge from KBs to improve word embedding models for biomedical NLP tasks. We propose a Graph regularized CBOW (GCBOW) model and a Graph regularized Skip-gram (GSkip-gram) model. GCBOW and GSkip-gram models use a graph to represent knowledge from KBs and integrate the graph regularization to basic CBOW and Skip-gram models respectively. The proposed models can be easily adapted to different types of KBs. In addition, we apply two different distance metrics for the graph regularization framework. Inspired by the contradictory results of applying word embedding to different tasks discussed in [178], we conduct experiments on both general domain tasks for intrinsic evaluation and biomedical NLP tasks for extrinsic evaluation. We evaluate our models on general word similarity datasets:

TOEFL synonym dataset, WordSimilarity-203, RG65, and SimLex-999. The results show that our models achieve promising improvement in precision on TOEFL synonym dataset and spearman's $\rho$ score on other three datasets. Furthermore, we evaluate on two biomedical NLP tasks: biomedical concept similarity/relatedness task and biomedical Information Retrieval (IR) task. Our method achieves better performances than baselines on both tasks.

Our major discoveries in this work are summarized as below:

- Integrating extra knowledge can improve the performance of word embedding models. The experiments on both general domain datasets and biomedical NLP tasks provide substantial evidence.
- For biomedical concepts similarity and relatedness tasks, GCBOW and GSkip-gram models achieve better results than baseline methods.
- Word embedding models improve biomedical IR task through concept weighting process. Bring extra knowledge from KBs improve the results.

The rest of this chapter is organized as follows. Section 5.2 describes the general word embedding models. Section 5.3 presents our knowledge graph representation, proposes graph regularized CBOW model and Skip-gram model, and develops the parameter updating for new proposed models. Section 5.4 describes our experimental results from intrinsic evaluation on standard datasets. Section 5.5 describes our experimental results from extrinsic evaluation on biomedical NLP tasks, and finally, Section 5.6 concludes this chapter.

## 5.2 Word Embedding Models

Word embedding models learn distributed representations of words from a large amount of unlabeled text data. Each word is represented as a dense and low-dimensional vector, and semantically similar words are transformed into similar vector representations. We take the CBOW and Skip-gram models proposed by Mikolov et al. (2013a, 2013b) [85, 86] as the basis for our proposed graph regularization framework.

Both CBOW and Skip-gram models are three-layer neural networks, containing input, projection, and output layers. The CBOW model learns word embedding by using context words to predict the center word $w_t$, where the context words refer to the neighbouring words within a window size $c$ near the centre word in a sentence. Given a sequence of training words $w_1, w_2, \ldots, w_T$, the CBOW model has the following objective function:

$$J_1 = \max \frac{1}{T} \sum_{t=1}^{T} \log p(w_t | w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}) \tag{5.1}$$

The Skip-gram model predicts surrounding words $w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}$ given the current centre word $w_t$. This model has the following objective function:

$$J_2 = \max \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \tag{5.2}$$

The probability $p(w_t | w_{t+j})$ is calculated using a softmax function:

$$p(w_t | w_{t+j}) = \frac{\exp(v_t^{'T} v_{t+j})}{\sum_{n=1}^{N} \exp\left(v_n^{'T} v_{t+j}\right)} \tag{5.3}$$

$v_n$ and $v_n^{'}$ are the input and the output representation vectors of word $w_n$. $N$ is the total vocabulary size. The representation vectors $v_n$ are between the input layer and projection layer, and $v_n^{'}$ are between projection layer and the output layer.

In the CBOW model, the projection layer $h$ is the average value of input representation of context words.

$$h = \frac{1}{2c} \sum_{-c \leq j \leq c, j \neq 0} v_{t+j} \tag{5.4}$$

In the Skip-gram model, the projection layer $h$ is the same as the input representation of word $w_t$, which is $v_t$.

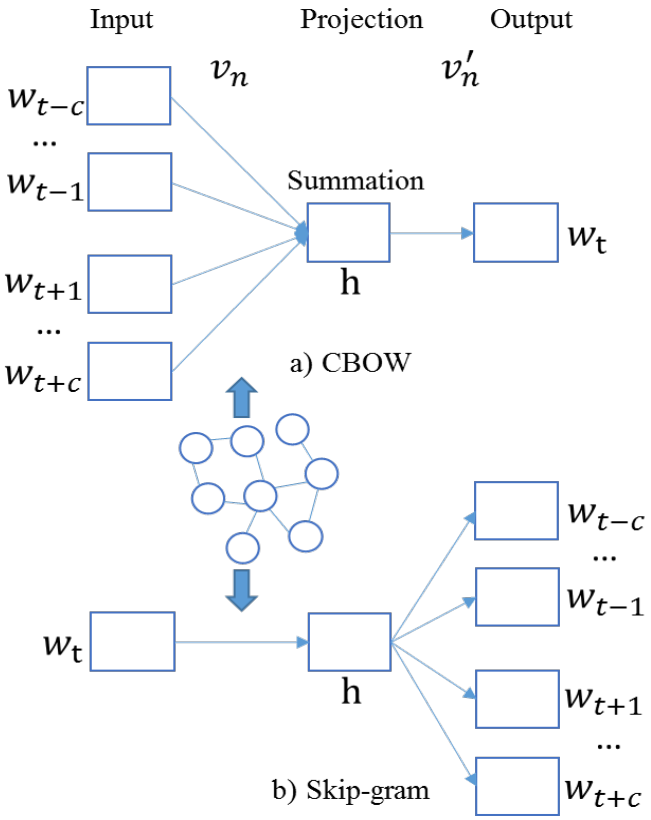



Figure 5.1 Word Embedding Models with Graph Regularization.

## 5.3 Graph Regularized Embedding Models

We use an undirected graph to represent knowledge from structured KBs. Relations between words from KBs can be represented as weighted edges between word nodes in the graph. We assume embedding representations of two words should be able to represent their closeness mentioned in KBs. We keep the assumption by adding a graph regularizer to the original objective function for CBOW model and Skip-gram model. The proposed graph regularization framework can use different distance metrics between words. In this study, we explore two specific distance metrics to build the graph regularizer.

### 5.3.1 Knowledge Graph Representation

The undirected graph as displayed in Fig. 5.1 represents relationships among words from extra knowledge sources. Each word is represented as a node in the graph. An edge connects nodes $n_i$

and $n_j$ if there is a relation mentioned in KBs between two nodes. A weight value is set for each edge connected between nodes $n_i$ and $n_j$. Different types of commonly used weighting schemas are discussed in the literature [180] [105]. We use a simple method to determine the weight value.

If two nodes $n_i$ and $n_j$ are connected because they are mentioned in KBs with similar meanings (e.g. synonym), we set the weight value $\omega_{ij} = 1$; if they are connected with opposite meanings, we set $\omega_{ij} = -1$; if they are connected with weak similar meanings, we set $\omega_{ij} = 0.5$. Here we define weak similar meanings as two words are related but not exactly have similar meanings. For example, in WordNet, if two words are indicated as hypernym or hyponym, we assume they have weak similar meanings.

### 5.3.2 Graph Regularization Framework

The embedding representations of two words represent their semantic relationships. Structured KBs enhance the representation of semantic information with graph structures. Thus we introduce graph regularized CBOW and Skip-gram model for incorporating the extra knowledge. Suppose word $w_i$ have relations with a set of other words $w_r, r \in \{1, \ldots, R\}$ in KBs. In our study, we use two types of distance metrics to measure the distance between two words $w_i$ and $w_j$. Here, $v_i$ and $v_j$ are vector representations for word $w_i$ and word $w_j$.

(1) Euclidean distance.

$$D_1(w_i, w_j) = ||v_i - v_j||_2 \tag{5.5}$$

(2) Divergence.

$$\begin{aligned} D_2(w_i, w_j) &= \frac{1}{2}(D(w_i||w_j) + D(w_j||w_i)) \\ &= \frac{1}{2}(\sum_{k=1}^{K} v_{ik} \log \frac{v_{ik}}{v_{jk}} + \sum_{k=1}^{K} v_{jk} \log \frac{v_{jk}}{v_{ik}}) \end{aligned} \tag{5.6}$$

$\omega_{ij}$ stands for the weight value between word node $w_i$ and $w_j$ (discussed in Section 5.3.1). By minimizing $\omega_{ij} D(w_i, w_j)$, we expect if two words have a close relation in KBs, their vector representations will also be close to each other. By adding this regularizer, we extend the original CBOW

model and Skip-gram model to the proposed GCBOW and GSkip-gram models. The GCBOW model has the following objective function:

$$J_3 = \max \frac{1}{T}\sum_{t=1}^{T}(1-\lambda)\log p(w_t|w_{t-c},...,w_{t-1},w_{t+1},...,w_{t+c}) - \lambda \sum_{-c\leq j\leq c, j\neq 0}\sum_{r=1}^{R}\omega_{t+j,r}D(w_{t+j},w_r) \tag{5.7}$$

$\lambda$ is a parameter to leverage the weights between the original objective and the newly added regularizer.

The GSkip-gram model has a similar objection function:

$$J_4 = \max \frac{1}{T}\sum_{t=1}^{T}(1-\lambda)\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) - \lambda\sum_{r=1}^{R}\omega_{tr}D(w_t,w_r) \tag{5.8}$$

### 5.3.3 Parameters Updating

We use stochastic gradient descent (SGD) to maximize the objective function for the GCBOW model and GSkip-gram model.

For the representation from projection layer to the output layer, hierarchical softmax is applied [85, 88]. Vocabulary is represented as a Huffmann binary tree. Each word $w_t$ can be reached by a path from the root of the tree. Let $L(w_o)$ be the length of the path. $n(w_o, j)$ is the j-th unit on the path from root to word $w_o$, and each unit has an output vector $v'_{n(w_o,j)}$. $ch(n)$ is an arbitrary fixed child of $n$. $[\![x]\!] = 1$ if $x$ is true, otherwise, $[\![x]\!] = -1$. In this path, each branch is treated as one binary classification. So $p(w_o|w_I)$ can be defined as follows:

$$p(w_o|w_I) = \prod_{j=1}^{L(w_o)-1}\sigma([\![n(w_o,j+1) = ch(n(w_o,j))]\!]v'^{T}_{n(w_o,j)}h) \tag{5.9}$$

For one training sample $\{w_i, w_o\}$, the training objective is $J = \max \log p(w_o|w_i)$ .

By taking the derivative of $J$ with regard to $v'_{n(w_o,j)}$ , we obtain

$$\begin{aligned}\frac{\partial J}{\partial v'_{n(w_o,j)}} &= \frac{\partial J}{\partial (v'_{n(w_o,j)})h}\frac{\partial (v'_{n(w_o,j)})h}{\partial v'_{n(w_o,j)}} \\ &= (t_j - \frac{1}{1+\exp(-v'_{n(w_o,j)}h)})h\end{aligned} \tag{5.10}$$

$t_j = 1$, if $n(w_o, j+1) = ch(n(w_o, j))$, otherwise, $t_j = 0$.

The update equation for representation from projection layer to output layer:

$$v'^{(new)}_{n(w_o,j)} = v'^{(old)}_{n(w_o,j)} + \alpha \frac{\partial J}{\partial v'_{n(w_o,j)}} \tag{5.11}$$

To learn the weights from the input layer to projection layer, we take the derivative of $J$ with regard to $v_i$:

$$\begin{aligned} \frac{\partial J}{\partial v_i} &= \sum_{j=1}^{L(w_o)-1} \frac{\partial J}{\partial (v'_{n(w_o,j)} h)} \frac{\partial (v'_{n(w_o,j)} h)}{\partial h} \frac{\partial h}{\partial v_i} \\ &= \sum_{j=1}^{L(w_o)-1} (t_j - \frac{1}{1+\exp(-v'^T_{n(w_o,j)} h)}) v'_{n(w_o,j)} \frac{\partial h}{\partial v_i} \end{aligned} \tag{5.12}$$

After adding the graph regularizer, we also need to take the derivative $A = \sum_{r=1}^{R} (\omega_{ir} D(w_i, w_r))$ with regard to $v_i$:

$$\frac{\partial A}{\partial v_i} = \frac{\partial \sum_{r=1}^{R} \omega_{ir} D(w_i, w_r)}{\partial v_i} \tag{5.13}$$

When using $D_1$ distance,

$$\begin{aligned} \frac{\partial A_1}{\partial v_i} &= \frac{\partial (\sum_{r=1}^{R} \omega_{ir} D_1(w_i, w_j))}{\partial v_i} \\ &= \sum_{r=1}^{R} \omega_{ir} (v_i - v_r) \end{aligned} \tag{5.14}$$

When using $D_2$ distance,

$$\begin{aligned} \frac{\partial A_2}{\partial v_i} &= \frac{\partial (\sum_{r=1}^{R} \omega_{ir} D_2(w_i, w_j))}{\partial v_i} \\ &= \sum_{r=1}^{R} \omega_{ir} \frac{1}{2} (\log \frac{v_{ik}}{v_{rk}} - \frac{v_{rk}}{v_i k} + 1) \end{aligned} \tag{5.15}$$

The update equation for representation from the input layer to projection layer:

$$v_i^{(new)} = v_i^{(old)} + \alpha ((1-\lambda) \frac{\partial J}{\partial v_i} - \lambda \frac{\partial A}{\partial v_i}) \tag{5.16}$$

## 5.4 Intrinsic evaluation

We conduct thorough experiments on four standard datasets to examine whether adding graph regularization can improve the performance of word embedding models. In this intrinsic evaluation, we explore different parameters settings of vector dimension size, windows size for context words,

$\lambda$ value, distance metrics. We also examine a few examples to discuss how the models improved by using extra knowledge from KBs. The goal of intrinsic evaluation is to evaluate our model on standard tasks for word embedding models and find the best parameters for the following biomedical NLP tasks.

### 5.4.1 Training Data

We train the word embedding models on the New York Times (NYT) corpus[1]. The dataset is pre-processed by sentence splitting, word tokenization, and stop words removal. We randomly sample 3M sentences from this corpus. The final training corpus contains 39,281,610 total words, and the unique words size is 268,032.

We use WordNet as KB and select three types of word pairs: Similar, Antonym and Hypernym. There are 106,828-word pairs in total.

### 5.4.2 TOEFL Synonym Selection Task

TOEFL synonym selection task [181] contains 80 target words, and the objective is to select the correct synonym for each target word from 4 candidate words. We get vector representations from embedding models for both target word and candidate words, and use the cosine similarity to calculate a score for each target word and candidate word pair, the one with the highest score is chosen as the final answer. The evaluation metric on this task is precision, which is the total number of questions with the correct answer divided by the total number of questions.

First, we use divergence ($D_2$) to evaluate the distance between two words. We chose different $d$ value and $\lambda$ value to compare GCBOW to CBOW, and GSkip-gram to Skip-gram. $d$ is the dimension size for word vector representation. We set the window size for context words $c = 5$.

Table 5.1 Performance (Precision, %) on TOEFL Synonym Dataset with $D_2$ Distance.

| | **CBOW** | **GCBOW** | | | **Skip-gram** | **GSkip-gram** | | |
|---|---|---|---|---|---|---|---|---|
| $d/\lambda$ | 0 | $5 \times 10^{-6}$ | $1 \times 10^{-5}$ | $5 \times 10^{-5}$ | 0 | $5 \times 10^{-6}$ | $1 \times 10^{-5}$ | $5 \times 10^{-5}$ |
| 50 | 51.9 | 54.4 | 50.6 | 51.9 | 53.8 | 53.8 | 57.5 | 53.8 |
| 100 | 54.4 | 60.8 | 59.5 | 57.0 | 63.8 | 66.3 | 52.5 | 63.8 |
| 200 | 58.2 | 64.6 | 62.0 | 60.8 | 66.3 | 68.8 | 70.0 | 63.8 |
| 300 | 58.8 | 60.0 | 60.0 | 56.3 | 68.8 | 70.0 | 55.0 | 53.8 |

[1] https://catalog.ldc.upenn.edu/LDC2008T19

Table 5.1 shows the results when $\lambda = 5 \times 10^{-6}$ and $d$ from $d = 50$ to $d = 300$. The GCBOW model has better performance than CBOW. The GSkip-gram also has better or equal performance than Skip-gram model. When ($d = 50$, $\lambda = 1 \times 10^{-6}$) and ($d = 300$, $\lambda = 5 \times 10^{-5}$), GCBOW has worse performance than CBOW. When ($d = 100$, $\lambda = 1 \times 10^{-5}$) and ($d = 300$, $\lambda = 5 \times 10^{-5}$ or $\lambda = 5 \times 10^{-5}$ ), GSkip-gram has worse performance than Skip-gram. According to the result, we recommend to set $\lambda = 5 \times 10^{-6}$, and $d = 200$ for both GCBOW and GSkip-gram models.

By setting different window sizes of context words for models, we get comparison results as shown in Table 5.2. In this experiment, we use $D_2$ distance, $\lambda = 5 \times 10^{-6}$ , and $d = 200$. With varying windows size, GSkip-gram always has the best performance. With window sizes 3 and 5, GCBOW outperforms CBOW.

Table 5.2 Performance (Precision, %) on TOEFL Synonym Dataset with Different Window Sizes.

| **Window Size** | **CBOW** | **GCBOW** | **Skip-gram** | **GSkip-gram** |
|---|---|---|---|---|
| 3 | 57.50 | 63.75 | 73.75 | 75.00 |
| 5 | 58.20 | 64.60 | 66.30 | 68.80 |
| 7 | 65.82 | 62.03 | 65.00 | 67.50 |

We evaluate models'performance with Euclidean distance ($D_1$) compared to $D_2$, we set $\lambda = 5 \times 10^{-6}$, $d = 200$, and $c = 5$. In Table 5.3, the models with $D_2$ distances have better performance than models with $D_1$ distances. For GCBOW, $D_2$ distance outperforms $D_1$ distance by 5.1% while for GSkip-gram, $D_2$ distance outperforms $D_1$ distance by 0.4%.

Table 5.3 Performance (Precision, %) on TOEFL Synonym Dataset with $D_1$ and $D_2$ Distance.

| **CBOW** | **GCBOW** | | **Skip-gram** | **GSkip-gram** | |
|---|---|---|---|---|---|
| | $D_1$ | $D_2$ | | $D_1$ | $D_2$ |
| 58.2 | 59.5 | **64.6** | 66.3 | 68.4 | **68.8** |

### 5.4.3 WS203, RG65 and SimLex-999 Datasets

The second group of standard datasets we use is WordSimilarity-353 (WS353) test collection [182, 183], RG65 [184] and SimLex-999 [185] datasets. These three datasets contain English word pairs along with human-assigned similarity judgments. The datasets are frequently used for evaluating word representations. The WS353 dataset is split into two subsets [183], one for evaluating similarity,

and the other for evaluating relatedness. We use the similarity part for our experiments, which contains 203 pairs (WS203). SimLex-999 contains 999 concrete and abstract adjective, noun and verb pairs with rating scores. RG65 is a smaller set containing 65 pairs.

The evaluation metric on this task is to compare correlations (Spearman's $\rho$ scores) between the similarity scores given by our models and those rated by human. Spearman's $\rho$ score measures the strength of association between two ranked variables. As displayed in Table 5.4, GCBOW with distance $D_1$ outperforms CBOW. GSkip-gram with distance $D_2$ has better performance than the Skip-gram model.

Table 5.4 Performance (Spearman's $\rho$ scores).

| **Dataset** | **CBOW** | **GCBOW** | | **Skip-gram** | **GSkip-gram** | |
|---|---|---|---|---|---|---|
| | | $D_1$ | $D_2$ | | $D_1$ | $D_2$ |
| WS203 | 0.751 | **0.761** | 0.745 | 0.655 | **0.664** | 0.659 |
| RG65 | 0.460 | **0.493** | 0.466 | 0.548 | 0.457 | **0.670** |
| Sim999 | 0.222 | **0.242** | 0.234 | 0.273 | 0.273 | **0.274** |

### 5.4.4 Qualitative Analysis

We examine the results from TOEFL synonym selection task to try to understand how the GCBOW and GSkip-gram models improved the performance over the CBOW and Skip-gram models. We identified four pairs of question and correct answer, which were correctly identified by GCBOW or GSkip-gram model but missed by the original CBOW model or Skip-gram model. Our analysis showed that there were three possible reasons for the improvement on performance:

(1) Explicit relations mentioned in KBs for a question and correct answer pair:

For some question and correct answer pairs, there are direct relations mentioned in KBs between them, we assume that is the reason why our model, which integrates knowledge from KBs, can improve the performance. One example is "$furnish/supply$" pair, there is a relation chain for them that exists in KBs.

furnish→HYPERNYM←supply

(2) Implicit relations mentioned in KBs for the question and correct answer pairs:

Implicit relation means there are no direct relations mentioned in KBs for the question and correct answer pair, but there are indirect relations between them. One example is "*temperate/mild*":

temperate→SIMILAR←moderate→SIMILAR←mild

The other example is "*root/origin*" :

root→HYPERNYM←become→HYPERNYM ←changeOfstate→HYPERNYM←

beginning→HYPERNYM←origin

We believe these implicit relations in KBs have led to performance improvements of our model.

(3) No relations mentioned in KBs for the question and correct answer pairs:

In some cases, there are no explicit or implicit relations exist in KBs for the question and answer words, but our models still work better. We assume there might be some patterns discovered by the models, but it remains unclear for us by now.

## 5.5 Extrinsic evaluation

We adopt best parameter settings from Section 5.4, and conduct experiments on two biomedical NLP tasks for the extrinsic evaluation. We exam the quality of our models by applying them to biomedical concepts similarity/relatedness task and biomedical IR task, and compare our methods with baselines from these tasks.

### 5.5.1 Training Data

We gather a biomedical corpus from two data sources: PubMed articles[2] and Clinical Medicine related Wikipedia articles[3]. The corpus contains over 5M sentences. We pre-process the dataset by conducting sentence splitting, word tokenization, and stop words removal. The total number of tokens is 93,095,323.

UMLS [29] is developed by the US National Library of Medicine and is a repository of biomedical vocabularies. We use UMLS MRREL table as our KB. This table defines relationships between UMLS concepts. There over 1.6M word pairs are selected from related relation types, such as

[2]https://www.ncbi.nlm.nih.gov/pmc/
[3]https://en.wikipedia.org/wiki/Category:Clinical_medicine

disease-treatment, disease-prevention, disease-diagnosis, disease-finding, sign and symptom, causes, and etc.

### 5.5.2 Biomedical Concepts Similarity and Relatedness

We apply our models to biomedical concepts similarity and relatedness tasks [186]. There are two datasets: UMNSRS-Similarity is a set of 566 UMLS concept pairs manually rated for semantic similarity, and UMNSRS-Relatedness is a set of 588 UMLS concept pairs manually rated for semantic relatedness. We also use the Spearman's $\rho$ scores as evaluation metric on this task.

Table 5.5 Performance (Spearman's $\rho$ scores) for Biomedical Concepts Datasets.

| **UMNSRS Dataset** | **Similarity** | **Relatedness** |
|---|---|---|
| Baseline [177] | 0.652 | 0.601 |
| CBOW | 0.755 | 0.734 |
| GCBOW | **0.775** | **0.747** |
| Skip-gram | 0.805 | 0.798 |
| GSkip-gram | **0.817** | **0.807** |

The results are displayed in Table 5.5. For both datasets, GCBOW outperforms CBOW. Also, GSkip-gram has better performance than the Skip-gram model. Except for using the CBOW and Skip-gram as intrinsic baselines, we also use the best result from [177] as an extrinsic baseline. Even we have a smaller corpus size (93,095,323 tokens) than extrinsic baseline (2,721,808,542 tokens), our models achieve a better result for this biomedical concepts similarity and relatedness tasks.

### 5.5.3 Concept Weighting for Biomedical IR

We utilize word embedding models for biomedical IR task through concept weighting process and conduct experiments for TREC 2015 Clinical Decision Support (CDS) task [120]. The task contains a clinical narratives dataset, which contains 30 topics, each topic is medical case narratives that describe scenarios related to patients medical history, signs/symptoms, diagnoses, tests, and treatments. The goal of this task is to return a ranked list of the top 1,000 articles from a collection of biomedical literature that are relevant to the medical case narratives. The biomedical articles collection contains around 733,000 articles from the PubMed Central (PMC)[4].

[4]http://www.trec-cds.org/2015.html#documents

Word embedding models are involved with concept weighting process as indicated in following steps:

***Step 1***: Identify concepts from narratives. We use MetaMap [187] to identify UMLS concepts in the case narratives. In order to avoid noises in this step, we also manually identify concepts as a comparison.

***Step 2***: Obtain weights for each concept. For each concept, a vector representation is obtained from embedding model. Each concept is measured using cosine similarity with all other concepts in order to obtain an average score. We use the score as concept weight value applied to document retrieval. We assume the more important concept will have a higher average score. A baseline is set by setting a weight value of 1 for all concepts (designated as C-1).

***Step 3***: Retrieve relevant documents. The basic retrieval model used is BM25 [188], and we use the weighted concepts from *step 2* to boost the retrieval results.

The first baseline for comparison is the best performing method in TREC 2015 (designated as C-trec) [189]. The other baselines used are C-1, and corresponding results generated from CBOW and Skip-gram. The evaluation measure is precision at top 5 retrieved documents (P@5).

Table 5.6 Performance (P@5) for Biomedical IR.

| **Concept Type** | **MetaMap** | **Manual** |
|---|---|---|
| C-1 | 0.3033 | 0.3467 |
| CBOW | 0.3067 | 0.4200 |
| GCBOW | 0.3233 | 0.4233 |
| Skip-gram | 0.3633 | 0.4400 |
| GSkip-gram | 0.3733 | **0.4667** |
| C-trec | 0.4467 | |

According to the results in Table 5.6, GCBOW has better performance than CBOW, and GSkip-gram also has better performance than Skip-gram. GSkip-gram with manual concepts achieves the best performance, which is better than two baselines: C-1 and C-trec. Manually concept identification has better performance than using MetaMap, that means by simply using MetaMap to identify concepts from narratives will introduce some noises.

## 5.6 Conclusions

This chapter presents two graph regularized word embedding models: GCBOW and GSkip-gram, which take extra knowledge from KBs into consideration. Experiments on standard word similarity tasks demonstrated that our models outperform the original CBOW and Skip-gram model correspondingly. We adopted best parameters setting from standard datasets evaluation and applied the models to two biomedical NLP tasks. Experimental results showed that integrating extra knowledge improved the performance for these two biomedical NLP tasks. Our models achieved better results than baselines in these tasks.

# Chapter 6: Clinical Diagnostic Inferencing

This chapter presents a novel approach to a novel task of automatically inferring the most probable diagnosis from a given clinical narrative [190]. Structured Knowledge Bases (KBs) can be useful for such complex tasks but not sufficient. Hence, we leverage the vast amounts of unstructured text and integrate the text with structured KBs. The key innovative ideas include building a concept graph from both structured and unstructured sources and ranking diagnosis concepts using the enhanced word embedding vectors learned from integrated sources. Experiments on the TREC CDS and HumanDx datasets showed that our methods improved the results of clinical diagnostic inferencing.

## 6.1 Introduction

Clinical diagnosis inference is the problem of automatically inferring the most probable diagnosis from a given clinical narrative. Many health information retrieval tasks will greatly benefit from the accurate results of clinical diagnostic inferencing. For example, in recent Text REtrieval Conference (TREC) Clinical Decision Support track (CDS[1]), diagnosis inference from medical narratives improved the accuracy of retrieving relevant biomedical articles [120, 72, 20].

Solutions to the problem require significant amount of input from domain experts and a variety of sources [117, 3]. To address the complex inference tasks, researchers [113, 191, 192] have utilized structured KBs that store structured information about various entity types and relation triples. Many large-scale KBs have been constructed over the years, such as WordNet [100], Yago [123], Freebase [101], DBpedia [124], NELL [125], UMLS Metathesaurus [29] etc. However, using KBs alone for inference task [118] suffer from limitations of the KBs. These limitations include incompleteness of knowledge, sparsity, and fixed schemas [193, 194].

On the other hand, unstructured textual resources such as Wikipedia generally contain more information than structured KBs. As a supplementary knowledge to mitigate the limitations of structured KBs, unstructured text combined with structured KBs provides improved results for

[1] http://www.trec-cds.org/

related tasks, for example, clinical question answering [195]. For processing text, word embedding models (e.g. skip-gram model [86, 85]) can efficiently discover and represent the underlying patterns of unstructured text. Word embedding models represent words and their relationships as continuous vectors. To improve word embedding models, previous works have also successfully leveraged the structured KBs [103, 97, 104].

Motivated by the superior power of the integration of structured KBs and unstructured text, we propose a novel approach to clinical diagnostic inferencing. The novelty lies in the ways of integrating structured KBs with unstructured text. Experiments showed that our methods improved clinical diagnostic inferencing from different aspects (Section 6.5.5). Previous work on diagnosis inference from clinical narrative either formulating it as a medical literature retrieval task [25, 26] or solving it with multiclass algorithms in a supervised way [27]. To the best of our knowledge, there is no work on diagnoses inference from clinical narratives conducted in an unsupervised way. Thus, we build baselines for this novel task.

## 6.2 Overview of the Approach

Our approach includes four steps in general: 1) extracting source concepts, $q$, from clinical narratives, 2) iteratively identifying corresponding evidence concepts, $a$, from KBs and unstructured text, 3) representing both source and evidence concepts in a weighted graph via a regularizer-enhanced skip-gram model, and 4) ranking the relevant evidence concepts (i.e. diagnoses) based on their association with the source concepts, $S(q, a)$ (computed by weighted dot product of two vectors), to generate the final output. Figure 6.1 shows the overview using an illustrative example.

Given source concepts as input, we build an edge-weighted graph representing the connections among all the concepts by iteratively retrieving evidence concepts from both KBs and unstructured text. The weights of the edges represent the strengths of the relationships between concepts. Each concept is represented as a word embedding vector. We combine all the source concept vectors into a single vector representing a clinical scenario. Source concepts are differentiated according to the weighting scheme in Section 6.4.2. Evidence concepts are also represented as vectors and ranked according to their relevance to the source concepts. For each clinical case, we find the most probable

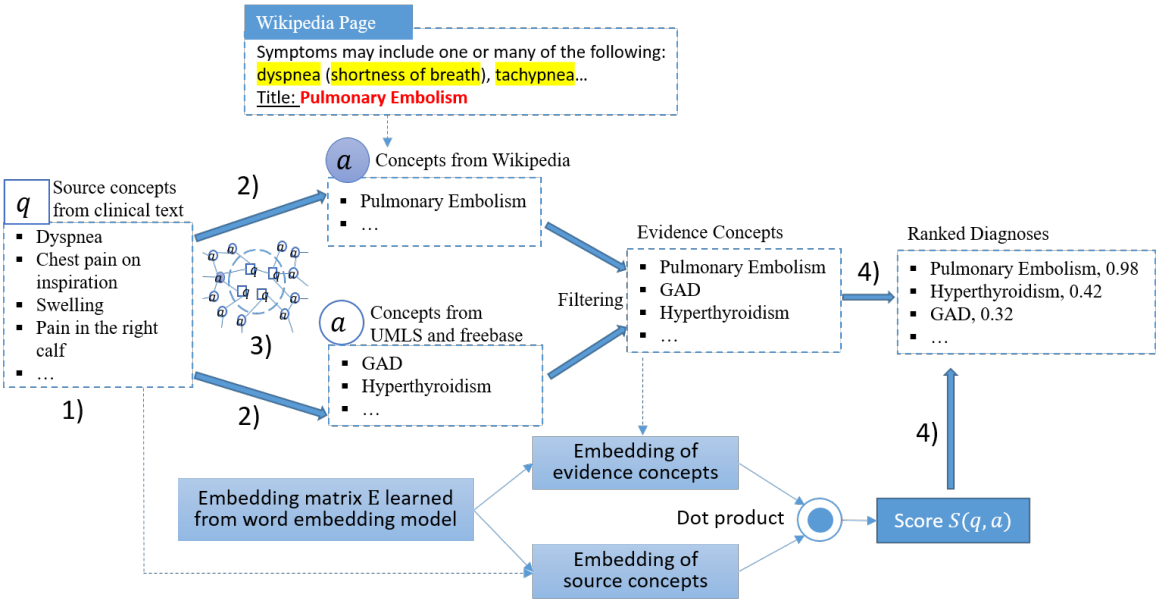



Figure 6.1 Overview of our system.

diagnoses from the top-ranked evidence concepts.

## 6.3 Knowledge Sources of Evidence Concepts

In this study, we use UMLS Metathesaurus [29] and Freebase [101] as the structured KBs. Both KBs provide semantic relation triples such as $\langle concept1, relation, concept2 \rangle$.

The Unified Medical Language System (UMLS [29])[2] is developed by the US National Library of Medicine and is a repository of biomedical vocabularies. The UMLS MRREL table defines the relationships between UMLS concepts. One example relation triple is $\langle concept : Giardiasis; relation : may\_be\_treated\_by; concept : Furazolidone \rangle$. We select UMLS relation types that are relevant to the problem of clinical diagnostic inferencing. These types include disease-treatment, disease-prevention, disease-finding, sign or symptom, causes etc. The details are displayed in Table 6.1.

Freebase [3] is a knowledge base contain a lot of triples from multiple domains, such as ⟨ subject; predicate; object ⟩. We select 61,243 triples from freebase that are related with medicine relation types. There are 19 such relation types in total. Most of them fall under the "medicine.disease" category, such as "causes", "risk factors", "symptoms" etc. One example relation triple in freebase is

[2]http://umlsks.nlm.nih.gov

Table 6.1 Selected Relation Types from UMLS MRREL.

| **Relation Category** | **Relation Type** |
|---|---|
| Disease-treatment | disease_has_accepted_treatment_with_regimen |
| | may_be_treated_by |
| | may_treat |
| | treated_by |
| | treats |
| Disease-prevention | may_be_prevented_by |
| | may_prevent |
| Disease-diagnosis | may_be_diagnosed_by |
| | may_diagnose |
| | diagnosed_by |
| | diagnoses |
| Disease-finding | disease_excludes_finding |
| | disease_has_finding |
| | associated_etiologic_finding_of |
| | associated_finding_of |
| | disease_may_have_finding |
| | has_associated_etiologic_finding |
| | has_associated_finding |
| | is_finding_of_disease |
| | may_be_finding_of_disease |
| Sign or symptom | has_sign_or_symptom |
| | sign_or_symptom_of |
| | has_manifestation |
| causes | cause_of |
| Associated disease | associated_disease |
| | disease_has_associated_disease |
| | disease_may_have_associated_disease |
| | is_associated_disease_of |
| | may_be_associated_disease_of_disease |
| Others | induces |
| | evaluation_of |

$\langle Giardiasis; medicine.disease.symptoms; Flatulence \rangle$. The 19 predicate types we choose are listed in Table 6.2.

Due to the incomplete of both KBs, we also use unstructured text as supplementary. For unstructured text, we use articles from Wikipedia and MayoClinic corpus as the supplementary knowledge source. Important clinical concepts mentioned in a Wikipedia/MayoClinic page can serve as a critical clue to a clinical diagnosis. For example, in Figure 6.1, we see that "dyspnea", "shortness of breath", "tachypnea" etc. are the related signs and symptoms of the "Pulmonary Embolism" diagnosis. We select 37,245 Wikipedia pages under the clinical diseases and medicine category in

Table 6.2 Selected Freebase Relation Types.

| | |
|---|---|
| 1 | medicine.condition_prevention_factors.conditions_this_may_prevent |
| 2 | medicine.diagnostic_test |
| 3 | medicine.disease_stage.stage_of |
| 4 | medicine.disease.causes |
| 5 | medicine.disease.includes_diseases |
| 6 | medicine.disease.parent_disease |
| 7 | medicine.disease.risk_factors |
| 8 | medicine.disease.symptoms |
| 9 | medicine.disease.treatments |
| 10 | medicine.drug_pregnancy_category.drugs_in_this_category |
| 11 | medicine.drug.drug_class |
| 12 | medicine.drug.mechanism_of_action |
| 13 | medicine.drug.route_of_administration |
| 14 | medicine.icd_9_cm_classification.includes_classifications |
| 15 | medicine.medical_specialty.diseases_treated |
| 16 | medicine.medical_treatment.used_to_treat |
| 17 | medicine.risk_factor.diseases |
| 18 | medicine.symptom.side_effect_of |
| 19 | medicine.symptom.symptom_of |

this study. In addition, MayoClinic[3] disease corpus contains 1,117 pages, which include sections of Symptoms, Causes, Risk Factors, Treatments and Drugs, Prevention, etc.

## 6.4 Methodology

### 6.4.1 Building Weighted Concept Graph

Both the source and the evidence concepts are represented as nodes in a graph. A clinical case is represented as a set of source concept nodes: $q = \{q_1, q_2, \ldots\}$. We build a weighted concept graph from source concepts using Algorithm 3.

Two kinds of evidence concept nodes are added to the graph: 1) the entities from KBs (UMLS and Freebase) (step 9-14 in Algorithm 3), and 2) the entities from unstructured text pages (step 15-20). If there exists a triple $< q_i, r, a_j >$ in KBs, where $r$ refers to a relation, an edge is used to connect node $q_i$ and node $a_j$. $w_{ij}$ represents the weight for that edge, and let $w_{ij} = 1$, if the corresponding triple occurs at least once. Due to the incompleteness of the KBs, there may exist multiple missing connections between a potential evidence concept $a_j$ and a source concept $q_i$. Unstructured knowledge from Wikipedia and MayoClinic can replenish these missing connections.

[3]http://www.mayoclinic.org/diseases-conditions

Algorithm 3 Build Concept Graph

1: **Input**: source concept nodes $q$
2: **Output**: graph $G = (V, E)$
3: $S = q$ and $V = q$;
4: **while** $S \neq \emptyset$ **do**
5: **for** each $q_i$ in $S$ **do**
6: **if** $distance(q_i, q) > 2$ **then**
7: continue;
8: **end if**
9: **if** triple $< q_i, r, a_j >$ in KBs **then**
10: $w_{ij} = 1$
11: $e = (q_i, a_j)$ and $e.value = w_{ij}$
12: insert $a_j$ to $V$ and $S$;
13: insert $e$ to $E$
14: **end if**
15: Use $q_i$ as query, search in Unstructured Text Corpora, get Result $R$
16: **for** each page-similarity pair $(p, s_{ij})$ in $R$ **do**
17: $e = (q_i, title(p))$ and $e.value = s_{ij}$;
18: insert $title(p)$ to $V$ and $S$;
19: insert $e$ to $E$;
20: **end for**
21: remove $q_i$ from $S$;
22: **end for**
23: **end while**

For each page $p$, the page title represents an evidence concept $a_j$. We use each source concept $q_i$ as a query, and page $p$ as a document, and then calculate a query-document similarity to measure the edge weight $w_{ij}$ between node $a_j$ and node $q_i$. We only take evidence concepts as all nodes connected to source concepts in a distance of at most 2 (step 6-8).

### 6.4.2 Representing Clinical Case

We combine the source concepts $q$ and get a single vector $v_q$ to represent the clinical narratives case. The source concepts from narratives for clinical diagnostic inferencing should be differentiated. Some source concepts are major symptoms for a diagnosis, while others are less critical. These major source concepts should be identified and given higher weight values. We develop two kinds of weighting schema for the differential expression of the source concepts. The source concept is represented as $v_q = \frac{1}{N} \sum_{q_i \in q} \gamma_i v_{q_i}$. $N$ is the total number of source concepts. $v_{q_i}$ is the vector representation for one source concept $q_i$.

(1) A longer concept usually convey more information (e.g. *malar rash* vs. *rash*), so it should be given more weights. We define this weight value as $\gamma_1 = \#Words\,in\,Concept$.

(2) For some commonly seen concepts (e.g. *fever*), usually, there are more edges connected to them. Sometimes, a common concept is less important for diagnosis inference, while some unique concepts are critical to infer a specific diagnosis. We define this weight value for each concept as $\gamma_2 = \frac{1}{\#Connected\,Edges}$. A higher weight value means the source concept is more unique.

### 6.4.3 Inferring Concepts for Diagnosis

**Extracting Potential Evidence Concepts:** From source concept nodes $q$, we find their connected concepts in the graph as evidence concepts. Traversing all edges in a graph is computationally expensive and often unnecessary for finding potential diagnoses. The solution is to use a subgraph. We follow the idea proposed in Bordes et al. [118]. The evidence concepts are defined as all nodes connected to source concepts in a distance of at most 2.

**Ranking Evidence Concepts:** We rank each evidence concept $a'$ according to its matching score $S(q, a')$ to the source concepts. The matching score $S(q, a')$ is a dot product of embedding representation of the evidence concept $a'$ and the source concept $q$ by taking the edge weights $w_{ij}$ into consideration. $S(q, a') = w_{ij} v_{a'} \cdot v_q$. $v_{a'}$ and $v_q$ are embedding representations for $a'$ and $q$. The embedding $E \in R^{k \times N}$ for concepts are trained using embedding models (Section 6.4.4). $N$ is the total number of concepts and $k$ is the predefined dimensions for the embedding vector. Each concept in the graph can find a $k$ dimensional vector representation in $E$. For a set of source concepts and evidence concepts $A(q)$, the top-ranked evidence concept can be computed as:

$$a = argmax_{(a' \in A(q))} S(q, a') \tag{6.1}$$

### 6.4.4 Word Embedding Models

We use the skip-gram model as the basic model (as discussed in section 2.4.2). The Skip-gram model predicts surrounding words $w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}$ given the current centre word $w_t$. We further enhance the skip-gram model by adding a graph regularizer. Given a sequence of training

words $w_1, w_2, \ldots, w_T$, the objective function is:

$$J = \max \frac{1}{T} \sum_{t=1}^{T} (1-\lambda) \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) - \lambda \sum_{r=1}^{R} D(v_t, v_r) \tag{6.2}$$

$v_t$ and $v_r$ are the representation vectors for word $w_t$ and word $w_r$. $\lambda$ is a parameter to leverage the graph regularizer and original objective. Suppose word $w_t$ is mentioned having relations with a set of other words $w_r, r \in \{1, \ldots, R\}$ in KBs. The graph regularizer $\lambda \sum_{r=1}^{R} D(v_t, v_r)$ integrates extra knowledge about semantic relationships among words within the graph structure. In our experiments, the distance between two concepts measured using KL-Divergence distance. $D(v_t, v_r)$ can be calculated using any other types of distance metrics. By minimizing $D(v_t, v_r)$, we expect if two concepts have a close relation in KBs, their vector representations will also be close to each other.

## 6.5 Experiments

### 6.5.1 Datasets for Clinical Diagnosis Inference

Our first dataset is from the 2015 TREC CDS track [120]. It contains 30 topics, where each topic is a medical case narrative that describes a patient scenario. Each case is associated with the ground truth diagnosis. We use MetaMap[4] to extract the source concepts from a narrative and then manually refine them to remove redundancy.

Our second dataset is curated from HumanDx[5], a project to foster integrating efforts to map health problems to their possible diagnoses. We curate diagnosis-findings relationships from HumanDx and create a dataset with 459 diagnosis-findings entries. Note that, the findings from this dataset are used as the given source concepts for a clinical scenario.

### 6.5.2 Training Data for Word Embeddings

We curate a biomedical corpus of around 5M sentences from two data sources: PubMed Central[6] from the 2015 TREC CDS snapshot[7] and Wikipedia articles under the "Clinical Medicine" category[8].

[4]https://metamap.nlm.nih.gov/
[5]https://www.humandx.org/
[6]https://www.ncbi.nlm.nih.gov/pmc/
[7]http://www.trec-cds.org/2015.html#documents
[8]https://en.wikipedia.org/wiki/Category:Clinical_medicine

After sentence splitting, word tokenization, and stop words removal, we train our word embedding models on this corpus. UMLS Metathesaurus and Freebase are used as KBs to train the graph regularizer. We use stochastic gradient descent (SGD) to maximize the objective function and set the parameters empirically.

### 6.5.3 Evaluation Metrics

We use Mean Reciprocal Rank (MRR) and Average Precision at 5 (P@5) to evaluate our models. MRR is a statistical measure to evaluate a process that generates a list of possible responses to a sample of queries, ordered by probability of correctness.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{6.3}$$

$|Q|$ is the total number of topics. $rank_i$ refers to the rank position of the correct diagnosis for the $i-th$ topic. The higher the MRR score, the better.

Average P@5 is calculated as precision at top 5 predicted results divided by the total number of topics. Since our dataset only has one correct diagnosis for each topic, all results have poor Average P@5 scores.

### 6.5.4 Results

Table 6.3 presents the results for our experiments. We report two baselines: *Skip-gram* refers to the basic word embedding model, and *Skip-gram** refers to the graph-regularized model using KBs. We also show the results for using different unstructured knowledge sources and different weighting schema. We can see that the best scores are obtained by the graph-regularized models with both the unstructured knowledge sources with variable weighting schema (Section 6.4.2).

### 6.5.5 Discussion

**Unstructured text is a critical supplement**: We analyze the source concepts and the corresponding evidence concepts for CDS topics and investigate the origin of the correct diagnoses. 70% of the correct diagnoses can be inferred from Wikipedia, 60% of the correct diagnoses from MayoClinic, 56% of the correct diagnoses from Freebase, and only 7% are from UMLS. Hence, Wikipedia and MayoClinic are very important sources for finding the correct diagnoses. The results indicate

Table 6.3 Evaluation results.

| | TREC CDS | | HumanDx | |
|---|---|---|---|---|
| **Method** | **MRR** | **Average P@5** | **MRR** | **Average P@5** |
| Baselines | | | | |
| Skip-gram | 21.66 | 8.88 | 18.56 | 5.08 |
| Skip-gram* | 22.60 | 8.88 | 18.63 | 5.15 |
| Skip-gram* + Different Unstructured Text Datasets | | | | |
| Wikipedia | 26.01 | 8.96 | 19.42 | 5.76 |
| MayoClinic | 32.64 | 9.52 | 19.46 | 5.80 |
| Both | 32.29 | 9.60 | 19.12 | 5.76 |
| Skip-gram* + Both Text Datasets + Different Weights | | | | |
| $\gamma_1$ | 32.22 | 10.40 | **21.09** | 5.88 |
| $\gamma_2$ | **32.77** | **12.00** | 20.86 | **5.93** |

that Freebase and UMLS are far from being complete, thus it is necessary to combine structured KBs with unstructured knowledge sources for clinical diagnostic inferencing.

**Source concepts should be differentiated**: In clinical narratives, some concepts are more critical than others for the clinical diagnostic inferencing. We developed two weighting schema to give more important concepts higher weight values. The results in Table 6.3 show that differentiating the source concepts with different weight values has a large impact on the model performance.

**Enhanced skip-gram is better**: We propose the enhanced skip-gram model by using a graph regularizer to integrate the semantic relationships among concepts from KBs. Experimental results show that diagnosis inference is improved by using word embedding representations from the enhanced skip-gram model.

## 6.6 Conclusion

We proposed a novel approach to a novel task of clinical diagnostic inferencing from clinical narratives. Our method overcomes the limitations of structured KBs by making use of the integrated structured and unstructured knowledge. The experimental results showed that the enhanced skip-gram model with differential expression of source concepts improved the performance on two benchmark datasets.

# Chapter 7: Conclusion and Future Work

## 7.1 Conclusion

Clinical text, such as clinical notes, contains lots of important information regarding a patient's medical conditions. Due to the limitations of lack of annotated clinical data, limited access to data, variation of clinical text, and limited extra knowledge sources, a systematic research is required to explore various methods and tools to better understand the clinical text. We gather data from clinical shared tasks and explore methods/tools to improve clinical text understanding from both corpus and document level.

In chapter 2, we summarize existing related work about clinical concept extraction, clinical document clustering, clinical relation extraction, word embedding models, and clinical diagnosis inference.

In chapter 3 and chapter 4, we focus on modeling different types of relationships existing in clinical notes. In chapter 3, we build a concept extraction system to extract medical concepts (e.g. symptoms, medications) from clinical text. Based on extracted clinical concepts, we apply a multi-view NMF method cluster clinical notes into meaningful groups. In chapter 4, we propose a Symptom-Medication (Symp-Med) matching framework to model symptom and medication relationships from clinical notes. After extracting symptom and medication concepts, we construct a weighted bipartite graph to represent the relationships between the two groups of concepts. We develop two Symp-Med matching algorithms to predict and recommend medications for symptoms.

In chapter 5, we focus on using extra knowledge from KBs to improve word embedding models for biomedical NLP tasks. We propose a Graph regularized CBOW (GCBOW) model and a Graph regularized Skip-gram (GSkip-gram) model. GCBOW and GSkip-gram models use a graph to represent knowledge from KBs and integrate the graph regularization to basic CBOW and Skip-gram models respectively. The proposed models can be easily adapted to different types of KBs. In addition, we apply two different distance metrics for the graph regularization framework. Our experimental

results on both standard datasets and biomedical NLP tasks show encouraging improvements with the new models.

In chapter 6, we present a novel approach to a novel task of automatically inferring the most probable diagnosis from a given clinical narrative. Structured KBs can be useful for such complex tasks but not sufficient. Hence, we leverage the vast amounts of unstructured text and integrate the text with structured KBs. The key innovative ideas include building a concept graph from both structured and unstructured sources and ranking diagnosis concepts using the enhanced word embedding vectors learned from integrated sources. Experiments on the TREC CDS and HumanDx datasets showed that our methods improved the results of clinical diagnosis inference.

## 7.2 Future Work

In this thesis, we propose methods for better understanding clinical text from both corpus and document level. More work needs to be done in the following directions:

(1) Integrating other types of clinical concepts and relationships in the graph.

In chapter 4, we build a Symp-Med weight matrix for our Symp-Med framework. We intend to extend it by using more clinical concept types, such as test, treatment, diagnosis, and etc. We also need to integrate other types of relationships, such as drug indications, side effects of drugs, drug interactions, drug administration information etc., from publicly available datasets such as DrugBank, RxNorm, and UMLS etc. [174].

(2) Improving concept extraction results using paraphrasing.

In chapter 3, we build a concept extraction system to extract medical concepts (e.g. symptoms, medications) from clinical text. We need to further improve the clinical concepts extraction accuracy from existing baselines. Currently, different systems may present a same clinical concept in different formats. For example, “hypertension” can be described as “high blood pressure” in one system, it can also be described as “HBP” in another system, we need to develop methods to be able to paraphrase such clinical concepts.

(3) Build intelligent diagnosis system.

In chapter 6, we present a novel approach to automatically infer the most probable diagnosis

from a given clinical narrative. In the future, we plan to extend to a complete intelligent diagnosis system. For given a list of symptoms describing a patient, we want to build a system capable of producing a correct diagnosis, treatment, test recommendations.

# Bibliography


[1] S. Henry, Pylypchuk and Patel, "Adoption of electronic health record systems among u.s. non-federal acute care hospitals: 2008-2015," 2016.

[2] W. Sun, A. Rumshisky, and O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 challenge," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 806–813, 2013.

[3] A. Lally, S. Bachi, M. A. Barborak, D. W. Buchanan, J. Chu-Carroll, D. A. Ferrucci, M. R. Glass, A. Kalyanpur, E. T. Mueller, J. W. Murdock *et al.*, "Watsonpaths: scenario-based question answering and inference over unstructured information," *Yorktown Heights: IBM Research*, 2014.

[4] Y. Ling, Y. An, M. Liu, and X. Hu, "An error detecting and tagging framework for reducing data entry errors in electronic medical records (emr) system," in *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. IEEE, 2013, pp. 249–254.

[5] D. R. Murphy, A. Laxmisan, B. A. Reis, E. J. Thomas, A. Esquivel, S. N. Forjuoh, R. Parikh, M. M. Khan, and H. Singh, "Electronic health record-based triggers to detect potential delays in cancer diagnosis," *BMJ quality & safety*, vol. 23, no. 1, pp. 8–16, 2014.

[6] D. R. Murphy, E. J. Thomas, A. N. Meyer, and H. Singh, "Development and validation of electronic health record–based triggers to detect delays in follow-up of abnormal lung imaging findings," *Radiology*, vol. 277, no. 1, pp. 81–87, 2015.

[7] K. El Emam, "Methods for the de-identification of electronic health records for genomic research," *Genome medicine*, vol. 3, no. 4, p. 1, 2011.

[8] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: a review of recent research," *BMC medical research methodology*, vol. 10, no. 1, p. 1, 2010.

[9] F. S. Collins and L. A. Tabak, "Nih plans to enhance reproducibility," *Nature*, vol. 505, no. 7485, p. 612, 2014.

[10] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner, "Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 540–543, 2011.

[11] N. Clynch and J. Kellett, "Medical documentation: Part of the solution, or part of the problem? a narrative review of the literature on the time spent on and value of medical documentation," *International journal of medical informatics*, vol. 84, no. 4, pp. 221–228, 2015.

[12] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of biomedical informatics*, vol. 42, no. 5, pp. 760–772, 2009.

[13] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle *et al.*, "Extracting information from textual documents in the electronic health record: a review of recent research," 2008.

[14] L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, L. Stoutenborough *et al.*, "Large-scale evaluation of automated clinical note de-identification and its impact on information extraction," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 84–94, 2013.

[15] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "Medex: a medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 19–24, 2010.

[16] 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," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[17] Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514–518, 2010.

[18] J. Patrick and M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 524–527, 2010.

[19] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh, "State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track," *Information Retrieval Journal*, vol. 19, no. 1-2, pp. 113–148, 2016.

[20] T. R. Goodwin and S. M. Harabagiu, "Medical question answering for clinical decision support," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 297–306.

[21] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[22] W. Boag, K. Wacome, and M. Tristan Naumann, "Cliner: A lightweight tool for clinical named entity recognition."

[23] M. Jiang, Y. Huang, J.-w. Fan, B. Tang, J. Denny, and H. Xu, "Parsing clinical text: how good are the state-of-the-art parsers?" *BMC medical informatics and decision making*, vol. 15, no. 1, p. 1, 2015.

[24] Y. Ling, Y. An, and X. Hu, "A matching framework for modeling symptom and medication relationships from clinical notes," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 515–520.

[25] Z. Zheng and X. Wan, "Graph-based multi-modality learning for clinical decision support," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 1945–1948.

[26] S. Balaneshin-kordan and A. Kotov, "Optimization method for weighting explicit and latent concepts in clinical decision support queries," in *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*. ACM, 2016, pp. 241–250.

[27] A. Prakash, S. Zhao, S. A. Hasan, V. Datla, K. Lee, A. Qadir, J. Liu, and O. Farri, "Condensed memory networks for clinical diagnostic inferencing," *arXiv preprint arXiv:1612.01848*, 2016.

[28] P. Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: a framework for mining clinical notes through symptom relation graphs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1167–1175.

[29] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[30] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, "Snomed clinical terms: overview of the development process and project status." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 662.

[31] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook *et al.*, "Loinc, a universal standard for identifying laboratory observations: a 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003.

[32] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, "Rxnorm: prescription for electronic drug information exchange," *IT professional*, vol. 7, no. 5, pp. 17–23, 2005.

[33] W. H. Organization *et al.*, "The icd-10 classification of mental and behavioural disorders: diagnostic criteria for research," 1993.

[34] A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD: NLM, NIH, DHHS*, pp. 1–26, 2006.

[35] J. D. Osborne, B. Gyawali, and T. Solorio, "Evaluation of ytex and metamap for clinical concept recognition," *arXiv preprint arXiv:1402.1668*, 2014.

[36] 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 medical informatics and decision making*, vol. 6, no. 1, p. 1, 2006.

[37] M. Jiang, Y. Wu, A. Shah, P. Priyanka, J. C. Denny, and H. Xu, "Extracting and standardizing medication information in clinical text–the medex-uima system," *AMIA Summits on Translational Science Proceedings*, vol. 2014, p. 37, 2014.

[38] J. Patrick, Y. Wang, and P. Budd, "An automated system for conversion of clinical notes into snomed clinical terminology," in *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*. Australian Computer Society, Inc., 2007, pp. 219–226.

[39] T. Hamon and N. Grabar, "Linguistic approach for identification of medication names and related information in clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 549–554, 2010.

[40] 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," *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301–310, 2001.

[41] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 646–651, 2010.

[42] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[43] W. H. Organization *et al.*, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization, 1992.

[44] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[45] ——, "A survey of text clustering algorithms," in *Mining text data*. Springer, 2012, pp. 77–128.

[46] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[47] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373–386, 2006.

[48] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative matrix factorization in multimodality data for segmentation and label prediction," in *16th Computer vision winter workshop*, 2011.

[49] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. of SDM*, vol. 13. SIAM, 2013, pp. 252–260.

[50] D. Hidru and A. Goldenberg, "Equinmf: Graph regularized multiview nonnegative matrix factorization," *arXiv preprint arXiv:1409.4018*, 2014.

[51] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: a survey," *Artificial intelligence in medicine*, vol. 33, no. 2, pp. 157–177, 2005.

[52] F. H. Saad, B. de la Iglesia, and D. G. Bell, "A comparison of two document clustering approaches for clustering medical documents." in *DMIN*, 2006, pp. 425–431.

[53] O. Patterson and J. F. Hurdle, "Document clustering of clinical narratives: a systematic study of clinical sublanguages," in *AMIA Annu Symp Proc*, vol. 2011. Citeseer, 2011, pp. 1099–1107.

[54] 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 7th international workshop on Data and text mining in biomedical informatics*. ACM, 2013, pp. 9–12.

[55] C. Han and J. Choi, "Effect of latent semantic indexing for clustering clinical documents," in *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*. IEEE, 2010, pp. 561–566.

[56] X. Zhang, L. Jing, X. Hu, M. Ng, J. Xia, and X. Zhou, "Medical document clustering using ontology-based term similarity measures," 2008.

[57] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations." in *ACL (1)*, 2014, pp. 402–412.

[58] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," 2013.

[59] J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier, "Connecting language and knowledge bases with embedding models for relation extraction," *arXiv preprint arXiv:1307.7973*, 2013.

[60] K. Toutanova, D. Chen, P. Pantel, P. Choudhury, and M. Gamon, "Representing text for joint embedding of text and knowledge bases," *ACL Association for Computational Linguistics*, 2015.

[61] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system." *Methods of information in medicine*, vol. 32, no. 4, pp. 281–291, 1993.

[62] C. Wang and J. Fan, "Medical relation extraction with manifold models." in *ACL (1)*, 2014, pp. 828–838.

[63] M. Hassan, O. Makkaoui, A. Coulet, and Y. Toussaint, "Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs," in *BioNLP 15*, 2015, p. 184.

[64] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," in *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2004, p. 430.

[65] H. Bruch and I. Hewlett, "Clinical notes: Psychologic aspects of the medical management of diabetes in children*." *Psychosomatic medicine*, vol. 9, no. 3, pp. 205–209, 1947.

[66] J. Main, R. Moss-Morris, R. Booth, A. A. Kaptein, and J. Kolbe, "The use of reliever medication in asthma: the role of negative mood and symptom reports," *Journal of Asthma*, vol. 40, no. 4, pp. 357–365, 2003.

[67] J. C. Slaughter, T. Lumley, L. Sheppard, J. Q. Koenig, and G. G. Shapiro, "Effects of ambient air pollution on symptom severity and medication use in children with asthma," *Annals of Allergy, Asthma & Immunology*, vol. 91, no. 4, pp. 346–353, 2003.

[68] R. P. Riechelmann, M. K. Krzyzanowska, A. OCarroll, and C. Zimmermann, "Symptom and medication profiles among cancer patients attending a palliative care clinic," *Supportive Care in Cancer*, vol. 15, no. 12, pp. 1407–1412, 2007.

[69] D. Häfner, K. Reich, P. Matricardi, H. Meyer, J. Kettner, and A. Narkus, "Prospective validation of allergy-control-scoretm: a novel symptom–medication score for clinical trials," *Allergy*, vol. 66, no. 5, pp. 629–636, 2011.

[70] J. E. Hopcroft and R. M. Karp, "An nˆ5/2 algorithm for maximum matchings in bipartite graphs," *SIAM Journal on computing*, vol. 2, no. 4, pp. 225–231, 1973.

[71] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood formation and anomaly detection in bipartite graphs," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 2005, pp. 8–pp.

[72] S. A. Hasan, Y. Ling, J. Liu, and O. Farri, "Using neural embeddings for diagnostic inferencing in clinical question answering," 2015.

[73] ——, "Exploiting neural embeddings for social media data analysis." in *TREC*, 2015.

[74] L. De Vine, M. Kholghi, G. Zuccon, L. Sitbon, and A. Nguyen, "Analysis of word embeddings and sequence features for clinical information extraction," 2015.

[75] T. Muneeb, S. K. Sahu, and A. Anand, "Evaluating distributed word representations for capturing semantics of biomedical concepts," *ACL-IJCNLP 2015*, p. 158, 2015.

[76] J. A. Minarro-Giménez, O. Marín-Alonso, and M. Samwald, "Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation," *arXiv preprint arXiv:1502.03682*, 2015.

[77] R. Socher, Y. Bengio, and C. Manning, "Deep learning for nlp," *Tutorial at Association of Computational Logistics (ACL), 2012, and North American Chapter of the Association of Computational Linguistics (NAACL)*, 2013.

[78] G. E. Hinton, "Distributed representations," 1984.

[79] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[80] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 641–648.

[81] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[82] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *Advances in neural information processing systems*, 2009, pp. 1081–1088.

[83] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[84] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 873–882.

[85] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[86] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[87] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[88] X. Rong, "word2vec parameter learning explained," *arXiv preprint arXiv:1411.2738*, 2014.

[89] Y. Goldberg, "A primer on neural network models for natural language processing," *arXiv preprint arXiv:1510.00726*, 2015.

[90] T. Luong, R. Socher, and C. D. Manning, "Better word representations with recursive neural networks for morphology." in *CoNLL*, 2013, pp. 104–113.

[91] M. Yu and M. Dredze, "Improving lexical embeddings with semantic knowledge." in *ACL (2)*, 2014, pp. 545–550.

[92] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," *arXiv preprint arXiv:1411.4166*, 2014.

[93] P. P. R. S. Asli Celikyilmaz, Dilek Hakkani-Tr, "Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems." AAAI - Association for the Advancement of Artificial Intelligence, January 2015.

[94] J. Cheng, Z. Wang, J.-R. Wen, J. Yan, and Z. Chen, "Contextual text understanding in distributional semantic space," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 133–142.

[95] W. Ling, C. Dyer, A. Black, and I. Trancoso, "Two/too simple adaptations of word2vec for syntax problems," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1299–1304.

[96] R. Cotterell and H. Schütze, "Morphological word-embeddings," in *Proc. of NAACL*, 2015.

[97] G. Zhou, T. He, J. Zhao, and P. Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering," in *Proceedings of ACL*, 2015, pp. 250–259.

[98] J. Bian, B. Gao, and T.-Y. Liu, "Knowledge-powered deep learning for word embedding," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 132–148.

[99] S. Virpioja, P. Smit, S.-A. Grönroos, M. Kurimo *et al.*, "Morfessor 2.0: Python implementation and extensions for morfessor baseline," 2013.

[100] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[101] 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*. ACM, 2008, pp. 1247–1250.

[102] C. Xu, Y. Bai, J. Bian, B. Gao, G. Wang, X. Liu, and T.-Y. Liu, "Rc-net: A general framework for incorporating knowledge into word representations," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 1219–1228.

[103] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph and text jointly embedding." in *EMNLP*. Citeseer, 2014, pp. 1591–1601.

[104] Q. Liu, H. Jiang, S. Wei, Z.-H. Ling, and Y. Hu, "Learning semantic word embeddings based on ordinal knowledge constraints," 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, pp. 1501–1511.

[105] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[106] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[107] E. Choi, M. T. Bahadori, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," *arXiv preprint arXiv:1511.05942*, 2015.

[108] 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 *AMIA Annu Symp Proc*, vol. 2011, 2011, pp. 171–180.

[109] J. Berant and P. Liang, "Semantic parsing via paraphrasing." in *ACL (1)*, 2014, pp. 1415–1425.

[110] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs." in *EMNLP*, vol. 2, no. 5, 2013, p. 6.

[111] Y. Shen, W. Rong, Z. Sun, Y. Ouyang, and Z. Xiong, "Question/answer matching for cqa system via combining lexical and sequential information." in *AAAI*, 2015, pp. 275–281.

[112] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1156–1165.

[113] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase." in *ACL (1)*. Citeseer, 2014, pp. 956–966.

[114] L. Heck and H. Huang, "Deep learning of knowledge graph embeddings for semantic parsing of twitter dialogs," in *Signal and Information Processing (GlobalSIP), 2014 IEEE global conference on*. IEEE, 2014, pp. 597–601.

[115] Y. Belinkov, M. Mohtarami, S. Cyphers, and J. Glass, "Vectorslu: A continuous word vector approach to answer selection in community question answering systems," *SemEval-2015*, p. 282, 2015.

[116] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 165–180.

[117] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: beyond jeopardy!" *Artificial Intelligence*, vol. 199, pp. 93–105, 2013.

[118] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," *arXiv preprint arXiv:1406.3676*, 2014.

[119] M. S. Simpson, E. M. Voorhees, and W. Hersh, "Overview of the trec 2014 clinical decision support track," DTIC Document, Tech. Rep., 2014.

[120] K. Roberts, M. S. Simpson, E. M. Voorhees, and W. R. Hersh, "Overview of the trec 2015 clinical decision support track." 2015.

[121] S. A. Hasan, X. Zhu, Y. Dong, J. Liu, and O. Farri, "A hybrid approach to clinical question answering," DTIC Document, Tech. Rep., 2014.

[122] M. Saeed, C. Lieu, G. Raber, and R. G. Mark, "Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring," in *Computers in Cardiology, 2002.* IEEE, 2002, pp. 641–644.

[123] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web.* ACM, 2007, pp. 697–706.

[124] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web.* Springer, 2007, pp. 722–735.

[125] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an architecture for never-ending language learning." in *AAAI*, vol. 5, 2010, p. 3.

[126] 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: Volume 2-Volume 2.* Association for Computational Linguistics, 2009, pp. 1003–1011.

[127] M. Liu, Y. Ling, Y. An, and X. Hu, "Relation extraction from biomedical literature with minimal supervision and grouping strategy," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on.* IEEE, 2014, pp. 444–449.

[128] V. Kumar, A. Stubbs, S. Shaw, and Ö. Uzuner, "Creation of a new longitudinal corpus of clinical narratives," *Journal of biomedical informatics*, vol. 58, pp. S6–S10, 2015.

[129] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[130] Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

[131] Ö. Uzuner, I. Goldstein, Y. Luo, and I. Kohane, "Identifying patient smoking status from medical discharge records," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 14–24, 2008.

[132] Ö. Uzuner, "Recognizing obesity and comorbidities in sparse data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561–570, 2009.

[133] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag, "Community annotation experiment for ground truth generation for the i2b2 medication challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 519–523, 2010.

[134] O. 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," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 786–791, 2012.

[135] Y. Xu, J. Liu, J. Wu, Y. Wang, Z. Tu, J.-T. Sun, J. Tsujii, I. Eric, and C. Chang, "A classification approach to coreference in discharge summaries: 2011 i2b2 challenge," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 897–905, 2012.

[136] W. Sun, A. Rumshisky, and O. Uzuner, "Annotating temporal information in clinical narratives," *Journal of biomedical informatics*, vol. 46, pp. S5–S12, 2013.

[137] A. Stubbs and Ö. Uzuner, "Annotating risk factors for heart disease in clinical narratives for diabetic patients," *Journal of biomedical informatics*, vol. 58, pp. S78–S91, 2015.

[138] W. F. Styler IV, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova *et al.*, "Temporal annotation in the clinical domain," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 143–154, 2014.

[139] S. Bethard, L. Derczynski, G. Savova, J. Pustejovsky, and M. Verhagen, "Semeval-2015 task 6: Clinical tempeval," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics Denver, Colorado, 2015, pp. 806–814.

[140] L. Deleger, T. Lingren, Y. Ni, M. Kaiser, L. Stoutenborough, K. Marsolo, M. Kouril, K. Molnar, and I. Solti, "Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research," *Journal of biomedical informatics*, vol. 50, pp. 173–183, 2014.

[141] E. M. Voorhees and W. R. Hersh, "Overview of the trec 2012 medical records track." in *TREC*, 2012.

[142] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, 2016.

[143] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Complex event extraction at pubmed scale," *Bioinformatics*, vol. 26, no. 12, pp. i382–i390, 2010.

[144] D. Demner-Fushman, J. G. Mork, S. E. Shooshan, and A. R. Aronson, "Umls content views appropriate for nlp processing of the biomedical literature vs. clinical text," *Journal of biomedical informatics*, vol. 43, no. 4, pp. 587–594, 2010.

[145] K. Denecke, "Extracting medical concepts from medical social media with clinical nlp tools: a qualitative study," in *Proceedings of the Fourth Workshop on Building and Evaluation Resources for Health and Biomedical Text Processing*, 2014.

[146] M. Xiaoyan Wang and M. Amy Chused, "Automated knowledge acquisition from clinical narrative reports," 2008.

[147] G. K. Savova, J. Fan, Z. Ye, S. P. Murphy, J. Zheng, C. G. Chute, and I. J. Kullo, "Discovering peripheral arterial disease cases from radiology notes using natural language processing," in *AMIA Annu Symp Proc*, vol. 2010, 2010, pp. 722–726.

[148] Y. Ling, X. Pan, G. Li, and X. Hu, "Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization," *IEEE transactions on nanobioscience*, vol. 14, no. 5, pp. 500–504, 2015.

[149] K. Roberts and S. M. Harabagiu, "A flexible framework for deriving assertions from electronic medical records," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 568–573, 2011.

[150] M.-Y. Kim, Y. Xu, O. Zaiane, and R. Goebel, "Patient information extraction in noisy telehealth texts," in *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. IEEE, 2013, pp. 326–329.

[151] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge *et al.*, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS Comput Biol*, vol. 7, no. 8, p. e1002141, 2011.

[152] G. Hripcsak, S. Bakken, P. D. Stetson, and V. L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery," *Journal of biomedical informatics*, vol. 36, no. 1, pp. 120–130, 2003.

[153] S. V. Pakhomov, A. Ruggieri, and C. G. Chute, "Maximum entropy modeling for mining patient medication status from free text." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2002, p. 587.

[154] A. Henriksson, "Semantic spaces of clinical text: leveraging distributional semantics for natural language processing of electronic health records," 2013.

[155] S. Kushinka, "Clinical documentation: Ehr deployment techniques," *California HealthCare Foundation*, 2010.

[156] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[157] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[158] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[159] Ö. Uzuner and A. Stubbs, "Practical applications for natural language processing in clinical research: The 2014 i2b2/uthealth shared tasks," *Journal of biomedical informatics*, vol. 58, pp. S1–S5, 2015.

[160] X. Huang, X. Zheng, W. Yuan, F. Wang, and S. Zhu, "Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization," *Information Sciences*, vol. 181, no. 11, pp. 2293–2302, 2011.

[161] C. for Disease Control, Prevention *et al.*, "National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011," *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*, vol. 201, no. 1, 2011.

[162] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner, "Identifying risk factors for heart disease over time: overview of 2014 i2b2/uthealth shared task track 2," *Journal of biomedical informatics*, vol. 58, pp. S67–S77, 2015.

[163] G. L. Booth, M. K. Kapral, K. Fung, and J. V. Tu, "Relation between age and cardiovascular disease in men and women with diabetes compared with non-diabetic people: a population-based retrospective cohort study," *The Lancet*, vol. 368, no. 9529, pp. 29–36, 2006.

[164] F. E. Kuhn and C. E. Rackley, "Coronary artery disease in women: risk factors, evaluation, treatment, and prevention," *Archives of internal medicine*, vol. 153, no. 23, pp. 2626–2636, 1993.

[165] J. S. Wilson, D. C. Shepherd, M. B. Rosenman, and A. N. Kho, "Identifying risk factors for healthcare-associated infections from electronic medical record home address data," *International journal of health geographics*, vol. 9, no. 1, p. 1, 2010.

[166] C. A. Harle, D. B. Neill, and R. Padman, "Information visualization for chronic disease risk assessment," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 81–85, 2012.

[167] X. Jiang, M. G. Langille, R. Y. Neches, M. Elliot, S. A. Levin, J. A. Eisen, J. S. Weitz, and J. Dushoff, "Functional biogeography of ocean microbes revealed through non-negative matrix factorization," *PloS one*, vol. 7, no. 9, p. e43866, 2012.

[168] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, "A comparative study of ontology based term similarity measures on pubmed document clustering," in *International Conference on Database Systems for Advanced Applications*. Springer, 2007, pp. 115–126.

[169] D. P. Williamson, "The primal-dual method for approximation algorithms," *Mathematical Programming*, vol. 91, no. 3, pp. 447–478, 2002.

[170] J. E. Mitchell, "Branch-and-cut algorithms for combinatorial optimization problems," *Handbook of applied optimization*, pp. 65–77, 2002.

[171] A. Makhorin, "Gnu linear programming kit," *Moscow Aviation Institute, Moscow, Russia*, vol. 38, 2001.

[172] D. P. Dobkin and S. P. Reiss, "The complexity of linear programming," *Theoretical Computer Science*, vol. 11, no. 1, pp. 1–18, 1980.

[173] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[174] M. Samwald, A. Jentzsch, C. Bouton, C. S. Kallesøe, E. Willighagen, J. Hajagos, M. S. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler *et al.*, "Linked open drug data for pharmaceutical research and development," *Journal of cheminformatics*, vol. 3, no. 1, p. 19, 2011.

[175] Y. Ling, Y. An, M. Liu, S. Hasan, Y. Fan, and X. Hu, "Integrating extra knowledge into word embedding models for biomedical nlp tasks," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017.

[176] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed research international*, vol. 2014, 2014.

[177] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, "How to train good word embeddings for biomedical nlp," *ACL 2016*, p. 166, 2016.

[178] B. Chiu, A. Korhonen, and S. Pyysalo, "Intrinsic evaluation of word vectors fails to predict extrinsic performance," *ACL 2016*, p. 1, 2016.

[179] P. Stenetorp, H. Soyer, S. Pyysalo, S. Ananiadou, and T. Chikayama, "Size (and domain) matters: Evaluating semantic word space representations for biomedical text."

[180] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." in *NIPS*, vol. 14, 2001, pp. 585–591.

[181] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[182] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 406–414.

[183] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 19–27.

[184] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[185] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, 2016.

[186] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton, "Semantic similarity and relatedness between clinical terms: an experimental study," in *AMIA annual symposium proceedings*, vol. 2010. American Medical Informatics Association, 2010, p. 572.

[187] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[188] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[189] S. Balaneshin-kordan, A. Kotov, and R. Xisto, "Wsu-ir at trec 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources," in *Proceedings of the 2015 Text Retrieval Conference*, 2015.

[190] Y. Ling, Y. An, and S. Hasan, "Improving clinical diagnosis inference through integration of structured and unstructured knowledge," in *In Proceedings of the 1st EACL 2017 Workshop on Sense, Concept and Entity Representations and their Applications (SENSE 2017)*, 2017.

[191] J. Bao, N. Duan, M. Zhou, and T. Zhao, "Knowledge-based question answering as machine translation," *Cell*, vol. 2, no. 6, 2014.

[192] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question answering over freebase with multi-column convolutional neural networks," in *Proceedings of Association for Computational Linguistics*, 2015, pp. 260–269.

[193] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in Neural Information Processing Systems*, 2013, pp. 926–934.

[194] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin, "Knowledge base completion via search-based question answering," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 515–526.

[195] A. Miller, A. Fisch, J. Dodge, A. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," *CoRR*, vol. abs/1606.03126, 2016.

# Vita

## Yuan Ling

### Education

- Drexel University, Philadelphia, Pennsylvania, USA
  - Ph.D., Information Science, March 2017
- Renmin University of China, Beijing, China
  - M.S., Management Science and Engineering, July 2013
- Beijing Jiaotong University, Beijing, China
  - B.S., Electronic Commerce, July 2010

### Selected Publications

- "Improving Clinical Diagnosis Inference through Integration of Structured and Unstructured Knowledge." *Accepted* by EACL 2017 Workshop on Sense, Concept and Entity Representations and their Applications, 2017.
- "Integrating Extra Knowledge into Word Embedding Models for Biomedical NLP Tasks." *To appear* In Proceedings of The International Joint Conference on Neural Networks (IJCNN), 2017.
- "Using Neural Embeddings for Diagnostic Inferencing in Clinical Question Answering." TREC, 2015.
- "Exploiting Neural Embeddings for Social Media Data Analysis." TREC, 2015.
- "Clinical documents clustering based on medication/symptom names using multi-view non-negative matrix factorization." IEEE Transactions on NanoBioscience, pages 500-504, 2015.
- "A Matching Framework for Modeling Symptom and Medication Relationships from Clinical Notes." In the IEEE Conference on Bioinformatics and Biomedicine (BIBM14), pages 515-520, 2014.
- "Relation Extraction from Biomedical Literature with Minimal Supervision and Grouping Strategy", In the IEEE Conference on Bioinformatics and Biomedicine (BIBM14), pages 444-449, 2014.
- "An Error Detecting and Tagging Framework for Reducing Data Entry Errors in Electronic Medical Records (EMR) System.", IEEE International Conference on Bioinformatics and Biomedicine (BIBM13), pages 249-254, 2013.

### Awards

- Best Student Paper Award at the IEEE International Conference on Bioinformatics and Biomedicine, 2013

### Public Outreach

- **NSF Center for Visual and Decision Informatics (CVDI)** *June 2013 to June 2015*
- **Philips Research North America** *May 2015 to December 2015*
- **Research Data Alliance/US Scholars Program** *June 2014 to August 2014*
- **TCL Research America** *June 2013 to September 2013*