Patient-Centered and Experience-Aware Mining for Effective Information Discovery in Health Forums

by

Yunzhong Liu

A Dissertation Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Approved May 2016 by the Graduate Supervisory Committee:

Yi Chen, Co-Chair
Huan Liu, Co-Chair
Baoxin Li
Hasan Davulcu

ARIZONA STATE UNIVERSITY
August 2016
ABSTRACT

Online health forums provide a convenient channel for patients, caregivers, and medical professionals to share their experience, support and encourage each other, and form health communities. The fast growing content in health forums provides a large repository for people to seek valuable information. A forum user can issue a keyword query to search health forums regarding to some specific questions, e.g., what treatments are effective for a disease symptom? A medical researcher can discover medical knowledge in a timely and large-scale fashion by automatically aggregating the latest evidences emerging in health forums.

This dissertation studies how to effectively discover information in health forums. Several challenges have been identified. First, the existing work relies on the syntactic information unit, such as a sentence, a post, or a thread, to bind different pieces of information in a forum. However, most of information discovery tasks should be based on the semantic information unit, a patient. For instance, given a keyword query that involves the relationship between a treatment and side effects, it is expected that the matched keywords refer to the same patient. In this work, patient-centered mining is proposed to mine patient semantic information units. In a patient information unit, the health information, such as diseases, symptoms, treatments, effects, and etc., is connected by the corresponding patient.

Second, the information published in health forums has varying degree of quality. Some information includes patient-reported personal health experience, while others can be hearsay. In this work, a context-aware experience extraction framework is proposed to mine patient-reported personal health experience, which can be used for evidence-based knowledge discovery or finding patients with similar experience.

At last, the proposed patient-centered and experience-aware mining framework is used to build a patient health information database for effectively discovering adverse
drug reactions (ADRs) from health forums. ADRs have become a serious health problem and even a leading cause of death in the United States. Health forums provide valuable evidences in a large scale and in a timely fashion through the active participation of patients, caregivers, and doctors. Empirical evaluation shows the effectiveness of the proposed approach.
To my family for their love and support
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Yi Chen, who has been an invaluable mentor guiding me in my research, and has contributed to this work in many ways. Her helps are always around, such as critical thinking, technical writing, and presentation skills. This dissertation would have never been possible without her support. I also want to thank my co-advisor, Dr. Huan Liu, who is a great mentor to follow. I benefit all my life from his insightful thoughts, words of wisdom, and great teaching style. I would also like to thank my committee members, Dr. Baoxin Li and Dr. Hasan Davulcu, who have provided valuable feedbacks, broadened my horizons, and helped me look beyond my current endeavors.

I would like to thank my colleagues at Arizona State University and New Jersey Institute of Technology: Yi Shan, Jinhe Shi, James Cotter, Mike Citro, and Viraj Bahalala, for their constructive criticism, helpful suggestions, and support. I would also want to thank members from the Data Mining and Machine Learning Lab: Dr. Jiliang Tang, Dr. Xia (Ben) Hu, Suhang Wang, Fred Morstatter, Dr. Shamanth Kumar, Liang Wu, Jundong Li, Tahora H. Nazer, and etc. They helped me a lot and greatly inspired me. I’m also grateful for the support from Google. Without the techniques I have learned during my two summer internships at Google, I would not be able to manage my research projects efficiently. In particular, I want to thank my two mentors at Google: Jian Lu and Griffin Deng.

There are many friends and professors who have encouraged and helped me during my PhD study. The first name goes to my friend Dr. Feng Chen, who persuaded me to come to the United States and has always been encouraging and ready to help. I would also like to thank Dr. Guoliang Xue, Weidong Zhang, Monica Dugan, Dr. Mutsumi Nakamura, Dr. Peng Zhang, Dr. Jun Shen, Qiongjie Tian, Lin Chen, Ruozhou Yu, and many others.
At last, I’m very grateful to my family for their love and understanding. Without their support, this thesis would not have been possible.

This dissertation is based upon work partially supported by NSF CAREER Award #1322406, Google, and Leir Charitable Foundations.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
</tbody>
</table>

## CHAPTER

1 INTRODUCTION ................................................. 1
  1.1 Background and Motivations ................................ 1
  1.2 Contributions ........................................... 8

2 PATIENT-CENTERED MINING ...................................... 10
  2.1 Patient-Centered Mining System ........................... 10
    2.1.1 Data Preprocessing .................................. 11
    2.1.2 Person Identification ................................ 12
    2.1.3 Person Resolution .................................... 12
    2.1.4 Patient Identification ................................ 13
    2.1.5 Health Information Association ....................... 14
  2.2 Person Resolution and Thread Structure Learning .......... 15
    2.2.1 Thread Structure Learning with Partially Labeled Data 19
    2.2.2 Leveraging Person Resolution for Thread Structure Learning 22
  2.3 Evaluation .................................................. 30
    2.3.1 Thread Structure Learning Evaluation .................. 30
    2.3.2 Patient-Centered Mining Evaluation .................... 37
  2.4 Summary ..................................................... 42

3 PATIENT EXPERIENCE MINING .................................. 43
  3.1 Introduction ............................................... 43
  3.2 Context Information for Patient Experience Mining ....... 44
  3.3 Context-Aware Experience Extraction ....................... 47
3.4 Experiments and Analysis .................................................. 50
  3.4.1 Experimental Settings ............................................... 50
  3.4.2 Baselines ............................................................. 51
  3.4.3 Performance Comparison ............................................ 53
  3.4.4 Impact of Context Information ..................................... 53
  3.4.5 Parameter Analysis .................................................. 54
3.5 Summary .............................................................................. 56

4 EFFECTIVE ADVERSE DRUG REACTION DISCOVERY .......... 57
  4.1 Introduction ........................................................................ 57
  4.2 Adverse Drug Reaction Discovery in Online Health Forums ...... 60
    4.2.1 Co-occurrence Analysis ............................................. 60
    4.2.2 Evaluation Methods .................................................. 61
    4.2.3 Term Mapping .......................................................... 61
  4.3 Experimental Evaluation .................................................. 62
    4.3.1 Data Set ................................................................. 62
    4.3.2 Ground Truth .......................................................... 63
    4.3.3 Comparison Methods ............................................... 64
    4.3.4 Evaluation with SIDER ............................................. 65
    4.3.5 Evaluation with Human Annotation ............................... 70
  4.4 Summary .............................................................................. 75

5 RELATED WORK ..................................................................... 76

6 CONCLUSIONS AND FUTURE WORK ............................... 82
  6.1 Conclusions ....................................................................... 82
  6.2 Future Work ....................................................................... 83
REFERENCES ................................................................. 86
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Comparison Methods</td>
<td>31</td>
</tr>
<tr>
<td>2.2 Performance Comparison on the Patients Forum Data Set</td>
<td>34</td>
</tr>
<tr>
<td>2.3 Performance Comparison on the Epilepsy Forum Data Set</td>
<td>35</td>
</tr>
<tr>
<td>2.4 Ten Randomly Chosen Queries</td>
<td>39</td>
</tr>
<tr>
<td>2.5 Evaluation for Ten Randomly Chosen Queries</td>
<td>41</td>
</tr>
<tr>
<td>3.1 Statistical Analysis of Label Pairs</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Performance Comparison</td>
<td>55</td>
</tr>
<tr>
<td>4.1 MedHelp Heart Disease Forum Data Set Statistics</td>
<td>63</td>
</tr>
<tr>
<td>4.2 Notation of the Comparison Methods</td>
<td>64</td>
</tr>
<tr>
<td>4.3 Evaluation with Human Annotation</td>
<td>72</td>
</tr>
<tr>
<td>4.4 False Positive Examples</td>
<td>73</td>
</tr>
<tr>
<td>4.5 False Negative Examples</td>
<td>73</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 An Overview of the Proposed Work</td>
<td>6</td>
</tr>
<tr>
<td>2.1 The Patient-Centered Mining System Overview</td>
<td>11</td>
</tr>
<tr>
<td>2.2 An Example of Semantic Role Labeling</td>
<td>14</td>
</tr>
<tr>
<td>2.3 A Sampled Thread of Eight Post Fragments</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Performance on Different Training Sizes (Patients Forum Data Set)</td>
<td>36</td>
</tr>
<tr>
<td>2.5 Performance on Different Training Sizes (Epilepsy Forum Data Set)</td>
<td>36</td>
</tr>
<tr>
<td>3.1 The Impact of of Context Information</td>
<td>55</td>
</tr>
<tr>
<td>3.2 The Impact of Local Context</td>
<td>56</td>
</tr>
<tr>
<td>4.1 Top-K Evaluation with SIDER for Aspirin among the Variants of the Proposed Method</td>
<td>66</td>
</tr>
<tr>
<td>4.2 Top-K Evaluation with SIDER for Atenolol among Variants of the Proposed Method</td>
<td>66</td>
</tr>
<tr>
<td>4.3 Top-K Evaluation with SIDER for Diazepam among Variants of the Proposed Method</td>
<td>67</td>
</tr>
<tr>
<td>4.4 Top-K Evaluation with SIDER for Aspirin by Comparing with the Existing Methods</td>
<td>68</td>
</tr>
<tr>
<td>4.5 Top-K Evaluation with SIDER for Atenolol by Comparing with the Existing Methods</td>
<td>68</td>
</tr>
<tr>
<td>4.6 Top-K Evaluation with SIDER for Diazepam by Comparing with the Existing Methods</td>
<td>69</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

1.1 Background and Motivations

Online health forums are a popular platform for people to share their personal experience, participate in discussions, express their feelings, and to support each other. There are a lot of online health forums available, such as MedHelp\(^1\), WebMD\(^2\), PatientsLikeMe\(^3\), Healthboards message boards\(^4\), and the Epilepsy forum\(^5\). The user population of such forums are rapidly growing. For instance, MedHelp currently has more than 13 million active monthly users. To achieve “smart health and well-being” demands patients to take an active role in understanding their health status and in making informed decisions. Health forums are an important resource for patients and caregivers to search for other patients with similar symptoms and to check what treatments have been taken by or suggested for those patients for self-education on their diseases and treatments. With highly valuable patient-contributed information and the ever increasing volume, health forums also provide the potential for doctors and medical researchers to discover knowledge about various diseases, treatments, their effects and adverse reactions, and so on. For instance, ADR has become a leading cause of death in the U.S.\(^{[12]}\). While traditional patient surveys or voluntary report systems\(^{[2]}\) are used to detect ADR, online health forums may

\(^{1}\)http://www.medhelp.org/
\(^{2}\)http://exchanges.webmd.com
\(^{3}\)http://www.patientslikeme.com
\(^{4}\)http://www.healthboards.com/boards
\(^{5}\)http://epilepsyfoundation.ning.com/forum
provide evidences in a much larger scale and in a timely fashion through the active participation of patients and caregivers.

However, there are some challenges that prevent us from effective information discovery in health forums.

**Patient-Centered Mining.** First of all, there is a mis-alignment between an existing syntactic information unit, e.g., a sentence, a post, or a thread in health forums, and a semantic information unit, e.g., a patient information unit, in which all the information is related to the same patient. Supposedly, most of information discovery tasks in health forums are based on patient information units, which are not available in health forums. In a health forum, a user, or a post author, may publish a post. An initial post and the replying posts submitted by the same or different authors compose a thread. While it is easy to share information by posts, and to browse and read the posts shared by other patients, current technology does not provide effective ways for a user to easily search information that she is interested in, in a large repository of posts. Let us look at two examples, both of which are observed in questions issued by real users to the epilepsy discussion forum.

Consider a user who wants to check other patients’ experience of using Vitamin for alleviating aggression in order to gain more knowledge. She would search information by issuing a multi-keywords query “Vitamin, aggression” on a health forum. One traditional approach, referred as post-based search in this work, returns a post as a search result if it contains all the input query keywords. Consider a scenario where a caregiver describes her daughter has “aggression” and seeks for suggestions, and another experienced forum user replies and suggests her daughter to take “Vitamin” without explicitly quoting the word “aggression”. While this would perfectly answer the user’s query, it will be missed by the posted-based search as the query keywords appear in two posts instead of one. Therefore, post-based search may suffer low recall.
To improve the recall, another approach, named as *thread-based search* in this work, can be adopted. It returns a thread or its web page link if all the posts in this thread collectively contain all the user input query keywords. Such an approach would be able to return the relevant result described earlier for the example query. However, suppose a user who suffers from seizures due to weaning wants to check other similar patients’ experience. She would issue a keyword query “seizure, wean”. Consider another thread discussing the effects of Vitamin B6 on an epilepsy patient. One post author mentions that her mother is taking Keppra to control her seizures while another caregiver mentions her son has weaned off Keppra since it causes anger. Although both of them have benefited from Vitamin B6, nobody has “seizure” related to “weaning”. This thread is thus not relevant to the user’s search intention. Such an irrelevant thread will be returned by the thread-based search as a query result, suffering low precision.

From the above examples, we observe that existing approaches do not perform well for a user query with multiple keywords. We analyze those queries and find that when a query contains multiple keywords, these keywords are expected to have close relationships between each other. For instance, a query may involve the relationship between a symptom and a disease, the relationship among several symptoms, the relationship among multiple diseases, the relationship between a disease and treatments, or the relationship between a treatment and side effects. To correctly find such relationships, it is critical that the matches to query keywords refer to the *same* patient. However, post-based or thread-based search does not consider *who* a keyword is associated with. They only check syntactic information units, either a post or a thread. It is common to see multiple posts refer to the same patient, and a thread contains information of multiple patients. Therefore the root cause of the low-quality results generated by existing approaches is the mis-alignment between the syntactic
information unit (a post or a thread) that the two traditional methods are based on and the semantic information unit (a patient) that the query user refers to.

In light of this observation, we propose to mine the semantic information unit - each individual patient and the associated information - from the posts [48, 13]. Then a user query is processed with respect to the semantic information unit, finding out the patients whose experience is related to query keywords and therefore can bring insights about the relationships among the keywords. We developed a patient-centered mining system, which takes the original forum data as the input, identifies the patients and the information associated with each individual, and outputs a patient-centered health information database. We also identified that the thread structure, the reply relationships between posts in a thread, is very important for patient-centered information mining. However, most online health forums only have partially labeled thread structures. We thus propose to learn the complete thread reply structures for better patient-centered mining [47, 49, 52], which will be introduced in Section 2.2.

**Patient Experience Mining.** Another challenge that prevents us from effective information discovery is varying degree of information quality in health forums. Some information includes patient-reported personal health experience, which is very valuable and provides real-world evidences for study. The other information can be just hearsay, which is not trustable and less valuable. To discover trustable information from health forums, it is critical for us to mine patient-reported personal health experience and differentiate it from the other forum information.

According to Wiki \(^6\), experience is a collection of events and/or activities from which an individual or group may gather knowledge, opinions and/or skills. In [63], experience is defined as knowledge embedded in a collection of activities or events which an individual or group has actually undergone. It can be subjective as in

\(^6\)http://en.wikipedia.org/wiki/Experience_(disambiguation)
opinions as well as objective. By following Wiki and [63], we define patient experience as the health-related events and/or activities undergone by a specific patient and the corresponding opinions on those events and/or activities perceived by that patient. For example, “Took Keppra for a week. I felt very nauseous” is patient experience. “Keppra could cause birth defects. It is really not a safe drug” can be part of patient experience if they are based on a past event that the patient has actually taken the drug and had the side effect. Otherwise, it can be just general knowledge, a guess, or hearsay. Note that, in health forums, a patient is sometimes represented by her/his caregiver, e.g., the patient’s mother. As they are closely related, the caregiver’s description or corresponding opinions are also counted as patient’s personal experience.

The patient experience mining is very important for effective information discovery in health forums. Lots of patients would like to seek other patients with similar experience other than general knowledge or suggestions from an expert. With those peers with similar experience, they can form a self-help group to support and encourage each other. In this case, patient experience extraction can help them find the right people to connect. Recently, the vast majority of research has been focused on online health forums to show its importance in knowledge discovery, e.g., discovering ADR knowledge [85, 46]. The online health forums provide a large-scale and timely fashion to obtain the ADR knowledge based on the aggregated evidences reported by patients or their caregivers. Some unverified hearsay, if used for ADR discovery, can be misleading. To discover trustable ADR knowledge from online health forums, it is also critical for us to mine patient-reported personal health experience.

**Effective Information Discovery.** To address the above challenges, in this work, patient-centered and experience-aware mining is proposed to build a patient health information database from the original health forum data. The overview of the pro-
The proposed work is shown in Fig. 1.1. The proposed framework can be used to build the foundation for many information discovery tasks, such as semantic information search or keyword query, evidence-based social media mining and knowledge discovery, and etc. Two types of potential applications are briefly introduced as follows.

First, it can provide effective query for forum users based on the mined patient information units [48]. Often a patient would like to know others who have similar conditions, symptoms, treatments, or adverse reactions, in order to gain more knowledge about the disease and to seek for social support. We can make recommendations on similar patients, and provide options for the query user to specify the query condition and what results they want. For example, they can select to retrieve either only the similar patients’ personal experience or their complete information, which can include the suggestions they received or their future treatment plan. If one patient has shared her information in the forum, we can extract her profile and use it as a query to find similar patients in the patient-centered and experience-aware health information database. We can automatically update the recommendation based on the latest information posted in the forum.

Figure 1.1: An Overview of the Proposed Work
Second, given a sufficient amount of data available and automatically extracted into our patient health information database, we can provide patient-centered and comprehensive medical information statistics to researchers. A simple application is to estimate the number of patients who have had a specific disease or taken a specific drug. Our method can be used to avoid counting the same patient’s experience multiple times as they were mentioned in different posts or threads. In addition to the relationship between drug treatments and adverse reactions, we can collect the statistical information about the relationships between population and symptoms/diseases, the relationships between symptoms and diseases, the relationships between multiple diseases, and etc. Such statistic information, summarized from individual patient cases, can bring new insights to medical research.

In this work, besides keyword query with patient-centered mining, we have mainly been focused on ADR discovery in health forums with patient-centered and experience-aware mining. Since the proposed framework connects the scattered health information by patients, it can identify some potential ADRs that are missed by the existing approaches. For example, if a patient took a drug a few weeks ago, and has a chronic adverse reaction now. Suppose the drug usage and adverse reaction information have been reported in different threads by the same patient or the patient’s caregiver. Our proposed approach can make the connection by aggregating the patient’s historical health information and identify the potential relationship between them. Our experience-aware mining approach can filter out those noisy information such as hearsay. For example, many patients may receive suggestions that the drug they are taking could cause some ADRs. Such suggested ADRs will not be counted with our experience-aware mining as long as no patients have directly experienced those ADRs in their reports. With the active participation of patients and caregivers, the proposed framework can discover ADRs in a large-scale and timely fashion. The doctors
or medical researchers can leverage our system to improve their ADR discovery process. They can retrieve the top-K potential ADRs returned by our system, and then further verify whether they are true or serious enough to take corresponding actions. With our patient health information database, they can also check the provenance of ADR evidences: the original context information that are related to each individual patient who reported the ADR experiences.

There are many challenges for the above keyword query or ADR discovery tasks. The first challenge is to obtain real-world data for the experiments. For example, we need to generate keyword queries similar to those issued by a real health forum user. Second, we need to design methods and evaluation metrics. For the ADR discovery task, we need to design how to aggregate the discovered evidences and rank them as potential ADRs for further evaluation. We evaluate the discovered potential ADRs with an official ADR knowledge base as well as human-annotated ground truth. However, there is a gap between the consumer terms used in health forums and the professional terms in the official ADR knowledge base. Ontology-based term clustering and mapping are required to fill such gap.

1.2 Contributions

The contributions of this research are summarized as follows:

- Proposing patient-centered mining to address the information mis-alignment challenge, which identifies each individual patient and associates the health information with the corresponding patient to form patient semantic information units for effective information discovery.

  - Identifying the importance of thread structures, the reply relationships between posts in a forum thread, for patient-centered mining. Proposing
to learn complete thread structures with partially known thread structures.

- Proposing a context-aware experience extraction framework to mine patient experience in health forums, which can be used to discover more reliable knowledge, or help forum users to search other patients with similar experience and connect with them.

- Proposing to use the patient-centered and experience-aware mining system to build a patient health information database for effective information discovery. Designing methods, evaluation metrics, and experiments for specific information discovery tasks in health forums. The empirical evaluation shows the effectiveness of the proposed approach.

The rest of this dissertation is structured as follows. Chapter 2 introduces patient-centered mining, including the patient-centered mining system and the thread structure learning framework. Chapter 3 introduces patient experience mining. Chapter 4 introduces ADR discovery with the patient-centered and experience-aware mining system. Chapter 5 surveys the related work. Chapter 6 concludes the completed work and discusses some future work.
Chapter 2

PATIENT-CENTERED MINING

2.1 Patient-Centered Mining System

To mine the patient health information units, we need to identify the patient mentions that refer to the same person and to associate and aggregate the health information with the corresponding patients. The patient-centered mining system overview is shown in Fig. 2.1, including the major components and data processing flow, the input and output for each component, and tools or techniques used by each component. The whole system takes the health forum data as the input and outputs the patient-centered health information units. In Fig. 2.1, the rectangle represents a function or data processing component, the diamond represents the intermediate input or output data, and the rounded rectangle represents the tool used by the corresponding component.

In addition to data collection and preprocessing, our patient-centered mining system also includes the following four major components: Person Identification, Person Resolution, Patient Identification, and Health Information Association. In Person Identification module, we discover all the person mentions to find the potential patient mentions. Since it is difficult to identify a patient from some individual person mentions, we apply Person Resolution to group all the person mentions into clusters such that all the mentions in the same cluster refer to the same person. Then we make Patient Identification based on all the information in each cluster of person mentions. At last, we make Health Information Association for each identified patient from the posts, and then output patient health information units into the patient-centered
Figure 2.1: The Patient-Centered Mining System Overview

health information database. In the following, we introduce each system component in details.

2.1.1 Data Preprocessing

In the data preprocessing component, we take the original forum data as input and output the post related information, thread structures, and separated sentences. The
information published in a health forum is usually organized in discussion threads. Each thread consists of a set of posts ordered by their posting time. In addition to the post content and posting time, each post is associated with an author, who wrote the post. We also extract the thread structure, the reply relationships between posts in a thread, which will be used in the following system modules. From each post content, we generate a sequence of sentences, which are the basic processing unit for most of the following tasks.

2.1.2 Person Identification

This component takes sentences in posts as input and outputs person mentions. Our method is based on the Stanford Core NLP tools [3] and MetaMap tool [6] in UMLS. First, all the person names identified by Named Entity Recognition (NER) and pronouns (except “it”) identified by Part of Speech (POS) tagger are identified as person mentions. For example, “Katie” (a person’s name), and some pronouns such as “she” and “her” are identified as person mentions. Second, we use the semantic types output by MetaMap to identify person mentions. For example, “my daughter” will be identified as a person mention since its semantic type is “family group”, which belongs to “living beings” semantic group.

2.1.3 Person Resolution

This component groups the person mentions within a thread into clusters such that each cluster includes all the mentions that refer to the same person. Stanford deterministic coreference resolution system [37] is used for generating person resolution within a post. For example, in “Katie is 5 years old, and she has a sensory integration disorder”, “Katie” and “she” are identified as co-referent.

In addition to person resolution within a post, we also do inter-post person res-
olution, which identifies the same person across posts and even across threads. We observed that in online health forums, many patients participated in multiple discussion threads to share their experience. They may mention that they took some drugs in a post, and then added some adverse reaction experience later in a follow-up post in the same thread or another thread with a similar topic. If we ignore such connections across posts or threads, we may miss the identification of some potential ADRs. Therefore, we proposed to connect these information pieces with the same patient by first identifying all the person mentions that refer to the same person across posts or threads.

As there are no publicly available systems for inter-post person resolution, we developed our own system based on some rules. We incorporate the author information and the thread structure for inter-post person resolution. First, we consider the same role associated with the same author refers to the same person. For example, if “my son” has been mentioned by the same author in two different posts, we consider them as co-referent. Second, we transform a thread into multiple multi-person conversation documents based on the reply relationship, in which a post author is a speaker and the post content is analogous to the utterance. In this way, within the same thread, the person mention in the replying post that refers to the person in its parent post can be identified as co-referent.

2.1.4 Patient Identification

This component identifies the patient mentions from the identified person mentions. We use semantic role labeling (SRL) [15] with Propbank [62] annotation to identify the semantic arguments associated with the predicate or verb in a sentence, and classify them into different semantic roles. In this work, we identify patients associated with some special verbs such as “diagnose”, “treat”, “cure”, “prescribe”,

13
“meditate”, and etc. One example generated by Illinois semantic role labeling visualization tool [66] is given in Fig. 2.2. In this example, “My daughter”, the $A_1$ argument of the verb “diagnose”, is identified as patient.

![Semantic Role Labeling Example](image)

**Figure 2.2:** An Example of Semantic Role Labeling

In addition, we also used 12 patient identification patterns based on a sample training data set, such as “take pharmacologic substance”, “have disease or syndrome”. Here “pharmacologic substance” and “disease or syndrome” are two semantic types for medical phrase, which can be extracted from post content by MetaMap. For example, from “she has a sensory integration disorder”, we can identify “she” is a patient since “sensory integration disorder” has the semantic type “disease or syndrome”. As shown in experimental evaluation later, this small number of patterns have a very high coverage in identifying patients and scale well in a large data set.

Since all the co-referent person mentions refer to the same person, if at least one of them has been identified as a patient mention, we identify this person as a patient.

2.1.5 Health Information Association

This component associates the health information in a forum thread with the corresponding patient mentioned in that thread. Here the health information can be the patient’s conditions, symptoms, treatments, diseases, adverse reactions, and
etc. They can be identified by MetaMap or matched given a keyword list. We then output patient health information units by aggregating the information associated with the same patient as one information unit. Note that the health information in a replying post can also be associated with the patient mentioned in its parent post if that replying post does not introduce a new patient. In this way, the clarification from a replier can also be associated with the patient mentioned in a preceding post.

2.2 Person Resolution and Thread Structure Learning

In the previous section, we have shown that the thread structure, the reply relationships between posts in a thread, is very important for patient-centered mining. The known post reply relationships have been used in the person resolution and information association components. However, most online health forums only have partially labeled structures. In this section, we propose to learn the complete thread reply structures.

A typical forum thread consists of a sequence of posts, ordered according to the time when the post is submitted. Logically, a thread can be represented by a tree structure, where each post has one parent to which it replies, except the first post, the root of the tree [79, 80, 77]. One post can be replied by multiple posts, that is, can have many children. An example of a forum thread in tree representation is shown in Fig. 2.3, which is extracted from one thread in the epilepsy foundation forum \(^1\). In this forum, if one post explicitly replies to another post, it will quote that post. We can easily obtain some explicit reply relationship from the quotation relationship, as indicated by the solid arrow in the figure. As we can see, Fig. 2.3 shows a partially labeled thread structure, where post 3, 4, and 5 do not quote any preceding post and thus have unknown parents.

\(^1\)http://epilepsyfoundation.ning.com/forum/topics/frontal-lobe-epilepsy-2
My son has frontal lobe as well as temporal lobe seizures. ...
I do hope that your son does well on Keppra and...

Our youngest son, now age 7 had a series of seizures at 18 months....

Wanted to start off by giving you some hugs. My son has been dealing with epilepsy since he was...

Hi, I just wanted to share my story a little. My son started out having...
... I wish u luck it is an extremely bumpy road.

@GM-I just read this thread now- I see you posted about your son having seizures in his sleep and you could tell ...
Thank You sooooo much for the hugs and also knowing that we are not alone....I would love to keep in touch and read about how your son and chat....

Thank you for your reply. I am sorry to hear the long road you have had with your son,...

It’s good to hear that your son has been doing well on Keppra- as is ours,...

The tree structure of forum threads can save users time and effort to track and get involved in the discussion, and help them to understand the interaction among forum users, such as who is following whom or who is the receiver of a suggestion. Literature also demonstrates that thread structure can boost the performance of automated forum information extraction [48], information retrieval [17, 78], clustering [64], online community search [69], topic summarization [53], and experts finding [90].

However, most of web forums do not have the complete thread structures available, which means the parents of some posts are unknown. Many forum authors just use the default mode to reply without specifying to which posts they reply, nor quoting existing posts. There is existing work for learning complete thread structures [7, 79, 77]. They require training data that have complete thread structures, which is typically obtained through labor-intensive manual labeling. We observed two proper-
ties in online health forums that we would like to leverage to learn thread structures in a scalable way without manually labeled training data. One is the prevalently available partially labeled thread structures in online forums, and the other is the key role that person references play in person-centric forums.

**Partially Labeled Thread Structures.** In reality, online forums have abundance of partially labeled reply structures. There are always some post authors who have a good habit of keeping an explicit reply structure. An example of such a partially labeled thread structure is shown in Fig. 2.3. In this forum, if one post explicitly replies to another post, it will quote that post. We can easily obtain some explicit reply relationships from the quotation relationship, as indicated by the solid arrow in the figure. While such partially labeled thread structures are prevalent in online forums, and can provide valuable information, they are not leveraged in existing work.

**Person-Centric Forums.** We observed two types of online forums. Some forums are centered around specific questions or topics, such as most of technical discussion forums. On the other hand, some forums are centered around persons, such as the health forums for patients and caregivers to share experience and support each other. Typically health forum users introduce problems and make comments in a subjective way, describing personal experience and giving feedbacks to other users. In other words, a thread has a collection of user cases raised by some forum users and commented by others. Since a post often refers to other persons either mentioned in this post, in the parent or ancestor post, identifying correct thread structure in person-centric forums is even more important than other forums to understand the context.

On the other hand, person-centric forums also bring opportunities. There are often person mentions in the posts. When one post replies to another, it tends to mention the person described in the parent or ancestor post. Conversely, if one post
mentions a person that is described in a preceding post, then this post is likely to be a child or descendant of that preceding post. According to this observation, if we can find out the person references, we can use them to help learning the thread reply structures. Indeed, often forums are written representation of conversations among a group of people. The references of persons provide great hints on who talks to whom in a “chat room”.

In this work, we first propose to learn complete thread structures from the partially labeled structures based on a statistical machine learning model: thread conditional random fields (threadCRF) [77]. Then we leverage the person reference information and combine it with threadCRF for thread structure learning. The person references can be obtained using person resolution techniques, which identify the same person mentioned in different context. We use unsupervised rule-based person resolution techniques to materialize the most likely candidates for unknown thread reply structures, and generate a fully labeled training data set. This data set can be considered as an approximation of the ground truth and used to bootstrap the supervised thread-CRF model training. We then use the learned model to re-label the unknown thread structures with the partially known structures as constraints. In addition to being used for training data generation, the person references are also encoded as semantic features and incorporated into the learning model to further improve the thread structure learning performance. By leveraging person references information discovered in semantic analysis of posts, and combining them with the syntactic and structural features captured by threadCRF, the proposed approaches provide a unified framework for thread structure learning. We have empirically verified the effectiveness of the proposed approaches.
2.2.1 Thread Structure Learning with Partially Labeled Data

In this section, we first define the thread structure learning problem and introduce the thread conditional random fields model. Then we introduce how to train and learn with the partially labeled thread structures.

**Problem Definition**

Given a thread $X_n$ with a sequence of $m$ posts $\{p_0, p_1, ..., p_{m-1}\}$, we need to find the parent post for each post in $X_n$, denoted by $Y_n = \{y_0, y_1, ..., y_{m-1}\}$, where $y_i$ is the known or predicted parent for $p_i$. Note that we only need to predict $y_i$ for $i > 1$, since the first post has no parent and the second post’s parent is always the first post.

**Thread Conditional Random Fields**

Thread conditional random fields (threadCRF), proposed by Wang et al. [77], is shown effective in learning thread structures. It is a supervised learning approach that requires a fully labeled data set for training. In threadCRF, given the post sequence in $X_n$ and the model parameter set $\Theta = \{\lambda_k\}_{k=1}^K$, the conditional distribution of $Y_n$ is defined as follows:

$$p(Y_n|X_n, \Theta) \propto \exp(\sum_{k=1}^K \lambda_k f_k(X_n, Y_n)),$$

(2.1)

where $\{f_k(X_n, Y_n)\}_{k=1}^K$ is the set of features for the post sequence in $X_n$ and the parent labeling sequence $Y_n$, and $\{\lambda_k\}_{k=1}^K$ are the weights for those corresponding features. The thread structure learning task is formulated as a Maximum a Posteriori (MAP) inference problem to find the optimal reply structure $Y^*$.

$$Y^* = \arg \max_{Y \in \Psi} p(Y|X_n, \Theta),$$

(2.2)

where $\Psi$ is the set of all possible reply structures for thread $X_n$. 

19
The key of the above threadCRF framework is to define a set of features to capture the interdependency among the posts in terms of the reply structure. 13 features are used, including 6 node features and 7 edge features. A node feature only depends on a pair of posts, say $p_i$ and $p_j$ with $i > j$, to determine how likely $p_i$ replies to $p_j$. For example, content similarity is one of such node features: If the content of $p_i$ is similar to that of $p_j$, $p_i$ is likely replying to $p_j$. An edge feature captures the dependency between two pairs of reply relationships. For example, one edge feature is repeat reply: If we know that Alice has replied to Bob, then the following post written by Bob is likely replying to Alice.

To handle the complexity associated with the edge features, which capture the long-distance dependency among posts, an approximate MAP inference is used for Eq. 2.2 to learn the model parameters from the training data set. Given a training set $T = \{X_1, X_2, ..., X_N\}$, with the ground-truth parent labels $R = \{Y_1, Y_2, ..., Y_N\}$, it estimates the optimal model parameters $\Theta = \{\lambda_k\}_{k=1}^K$ by maximizing the following log-likelihood function:

$$L_\Theta = \sum_{n=1}^{N} \log p(Y_n|X_n, \Theta)$$

$$= \sum_{n=1}^{N} [\Theta^T F(X_n, Y_n) - \log Z_\Theta(X_n)],$$

where $F(X_n, Y_n)$ are the accumulated feature values for one thread in the training set and $\log Z_\Theta(X_n) = \sum_Y \exp(\Theta^T F(X_n, Y))$. L-BFGS algorithm is used to optimize the object function in Eq.2.3. The gradient is derived by taking the derivative of the object function.

$$\nabla L_\Theta = \sum_{n=1}^{N} [F(X_n, Y_n) - E_{p_\Theta(Y|X_n)} F(X_n, Y)] - \frac{\lambda}{\sigma^2},$$

where $E_{p_\Theta(Y|X_n)} F(X_n, Y)$ is the model expectation of the features’ occurrences for the given training thread, and $\frac{\lambda}{\sigma^2}$ is the regularization term.
**Training Set Generation with Partially Labeled Data**

Note that threadCRF is a supervised learning model, which requires a completely labeled data set for model training. In this subsection, we propose to generate a fully labeled training set given the partially labeled data.

We materialize all the possible thread reply structures given the partially labeled reply structures. Specifically, if a post has explicitly specified its parent, then we use this information directly. Otherwise, we consider each preceding post as a possible parent of the post, and generate multiple possible training instances with each containing a possible thread reply structure. In this way, the obtained training data sets are fully labeled. We call this process of converting a partially labeled data set to a fully labeled one as *materialization*.

We consider all the materialized instances from the same partially labeled thread equally possible. We denote this approach as MEP, which means Materialization with Equal Probabilities. Assume that the \( n \)th partially labeled training thread can be materialized into \( M_n \) completely labeled instances. With the materialized training instances, we have the following equation for the derivative of the threadCRF object function, which is modified from Eq. 2.4.

\[
\nabla L_{\Theta} = \sum_{n=1}^{N} \left[ \frac{\sum_{i=1}^{M_n} F(X_n, Y_{n_i})}{M_n} - E_{\rho_{n}}(Y|X_n) F(X_n, Y) \right] - \frac{\lambda}{\sigma^2},
\]

where \( X_n \) is the post sequence of the \( n \)th training thread, and \( Y_{n_i}, 1 \leq i \leq M_n \), is one possible parent labeling sequence for the \( n \)th thread. \( \sum_{i=1}^{M_n} F(X_n, Y_{n_i}) \) is the accumulated empirical feature value for \( X_n \).

This materialization approach considers all the possible reply structures in a thread equally important, which may not be accurate. Furthermore, a huge amount of materialized training instances will be generated, which will lead to a dramatically increasing in the time and space complexity for the factor graph generation and
marginal probability inference during the threadCRF model learning. For example, if there are \( T \) posts, denoted as \( p_{u_1}, p_{u_2}, \ldots, p_{u_T} \), with unknown parents in a thread, where \( p_{u_i} \) has \( u_i \) candidate parents and \( u_i \geq 2 \). Then the total number of materialized instances will be \( u_1 \cdot u_2 \cdot \ldots \cdot u_T \geq T! \), which is a huge number if \( T \) is relatively large. A more effective and efficient materialization process will be introduced later.

**Constrained ThreadCRF for Partially Labeled Data**

When applying the trained model to learn a complete reply structure for a given thread, threadCRF predicts the parents for all the posts despite the fact that some of them are already known. We propose to use the existing partially known structure as constraints. We denote this approach as constrained threadCRF. We not only want to preserve the existing reply structures in the final output of the complete reply structures, but also want the existing structures to help infer the unknown structures by encoding them into the model. In order to do that, we add one parent feature into the original threadCRF model, which is defined as follows:

\[
Parent(y_i = j) = \begin{cases} 
1 & \text{if } y_i = j \text{ is known;} \\
-1 & \text{if } y_i \neq j \text{ is known;} \\
\frac{1}{i} & \text{if } y_i \text{ is unknown.}
\end{cases}
\]  

(2.6)

Here \( y_i \) is the parent post ID of the \( i \)th post, and \( j \) is a parent post ID, where \( i \geq 1 \), \( j \geq 0 \), and \( i > j \). For the \( i \)th post, if its parent is unknown, all the \( i \) candidates are assigned with the same feature value \( \frac{1}{i} \).

2.2.2 Leveraging Person Resolution for Thread Structure Learning

As discussed previously, person resolution can be useful for thread structure learning in health forums. In this section, we discuss how to leverage person resolution for thread structure learning in health forums. We first introduce our person resolution
system, then we present a rule-based system that uses person resolution to generate thread structures. At last, we introduce how to combine person resolution with threadCRF.

**Rule-based Person Resolution**

Person resolution (PR) is the process of identifying the same person mentioned in different contexts. Usually, some general coreference resolution, anaphora resolution, or pronoun resolution systems can be used for person resolution [38, 37, 73, 55, 74]. In terms of the scope, there are three types of person resolution in a forum: intra-post, inter-post, and inter-thread person resolution. The intra-post person resolution confines the person resolution within a post. The inter-post person resolution considers the person resolution between posts but within a thread, while the inter-thread person resolution considers the person resolution between threads.

We mainly use the inter-post person resolution for thread structure learning, since the thread reply structures focus on the relationships between posts within a thread. The intuition behind that is as follows: If one post replies to another post, it tends to mention the same person who has been mentioned in its parent post. Conversely, if one post contains some person mentions that refer to the same person mentioned in a preceding post, then the post is likely to be a child or a descendant of that preceding post.

We design our own inter-post person resolution system for thread structure learning mainly for two reasons. First, there are no publicly available systems for inter-post person resolution. Second, since our goal is to leverage person resolution for thread structure learning, we only need to do person resolution for some person mention pairs that are from two different posts. We observed that a forum thread can be considered as a multi-person dialogue, where a post author is like a speaker and the post
content is analogous to the utterance, though a post can be very long. We designed several types of PR features. Each feature type has a different priority. These feature types are arranged in the descending order of priority, which serve as a multi-pass sieve with the first pass (type) having the highest priority. Specifically, our current PR system for thread structure learning includes the following four types of features in the order of descending priority.

**PR Feature Type 1:** Matching between the addresses in the current post content and the signature in the parent post content. Usually, the address appears at the beginning or follows some token like “hi”, “hello”, and etc., while the signature appears at the end of a post following some tokens like “thanks”, “regards”, and etc. We not only consider a person name recognized by a name entity recognition (NER) system in Stanford Core NLP tools as an address or a signature, but also identify those nicknames, acronyms, authorID, and etc., based on some common patterns expressed by regular expressions.

**PR Feature Type 2:** Matching between the same role related to the same person. First, our system identifies all the role mentions, such as “son”, “daughter”, “sister”, and etc., using the family group semantic type in MetaMap [6]. Second, we combine the identified role with the first or second personal pronouns like “our”, “my”, and “your” for matching pairs, where the pronouns are identified by the Part-of-Speech (POS) module in Stanford Core NLP tools. Finally, the word “my/our” followed by a role (such as daughter) in a preceding post can match “your” followed by the same role in the current post.

**PR Feature Type 3:** Matching between the first person pronouns, like “I” and “we”, in the candidate parent post and the second person pronouns like “you” in the current post that tend to refer to the same person. Semantic role labeling (SRL) [15] and WordNet [19] are used for checking if they tend to refer to the same person. First,
we identify the first person pronouns in the parent, and the second person pronouns in the current post. Then we use SRL for finding the associated verbs with the pronouns in a sentence. If the verb associated with the first person pronoun is a synonym of the verb associated with the second person pronoun according to WordNet, then we consider them as a matching. For example, “I” in “I got fever” can be matched to “you” in “The symptom you had...”, as “get” and “have” are synonyms in WordNet.

**PR Feature Type 4:** Matching between the third person pronoun in the current post and the person name or role in the parent post that are consistent in gender. Note that we also ensure that the third person pronoun has not been resolved to a person name or role appearing in the same post, and this person is a different one from that in the parent post. The Stanford coreference resolution system is used for checking the person resolution within a post. For example, “she” in the first sentence of a post, “she should ... ”, can refer to “my daughter” mentioned in a proceeding post. But “she” in a post like “My friend Mary ... she ...” cannot refer to “my daughter” in a proceeding post, as here “she” refers to “Mary”, who is a different person.

Here we only define pairwise PR features, which means each feature only involves a pair of posts. When applying person resolution for the unsupervised thread structure generation, we predict the unknown thread structure by evaluating each individual reply relationship independently. Note that we can also define the PR features involving multiple pairs of posts corresponding to multiple reply relationships, and learn the entire thread structure tree by considering all these PR features at the same time. In this way, we learn a globally optimal thread structure in terms of person resolution. However, such a global optimization will be highly expensive in computation.
**Person Resolution for Thread Structure Generation**

Given the above different types of person resolution features, we design the multi-pass candidate structure selection algorithm as shown in Algorithm 1, which takes a thread with a partially known structure as input and outputs a complete thread structure. We compare the priority of feature types matched by each candidate. In particular, each type of features, in the descending order of priority, is used to filter out those less likely candidates. For each feature type, we divide the candidates into two subsets: matched and unmatched. We then remove the unmatched from the candidate set. We continue in this way until all the types (passes) of features have been checked.

We break a tie when there are multiple candidates left at the end using the following two rules. First, we observed that the forum users tend to reply to the thread initiator, who is the author of the first post, except for the initiator self. If a post not authored by the thread initiator has a set of candidate parents, and one of them is from the thread initiator, then we output that post as the labeled parent. Second, we observed that forum users tend to reply to the latest post given other factors the same. If two candidate parents are both from the thread initiator or neither of them is from the thread initiator, then the latest one will be output as the labeled parent.

**Combining PR with ThreadCRF**

In this subsection, we introduce how to combine person resolution with threadCRF, including how to materialize training instances with person resolution evaluation, and how to encode person resolution into threadCRF.

**Materialization with PR Evaluation.** As we have discussed earlier, materializing partially labeled data to fully labeled one with equal probabilities will result in an
Algorithm 1: Multi-Pass Candidate Thread Structure Selection

Input: One thread with a partially known structure.

Output: A complete thread structure.

1 for each post do
2    if its parent post is unknown then
3        Put all the preceding posts in the candidate set;
4        for each type of PR features (in the descending order of priority) do
5            for each parent in the candidate set do
6                Check if this type of features can be matched;
7            end
8            if Only one candidate has the matched feature then
9                Output that candidate as the labeled parent;
10           else if More than one matched candidates then
11                Remove all the other unmatched candidates from the candidate set;
12            end
13        end
14    if More than one candidate left then
15        Break the tie according to the two defined rules and output the winner as the labeled parent;
16    else
17        Output the known parent.
18 end
exponential increase on the data size and thus the computational time and space in model training process. Furthermore, not all preceding posts are equally likely to be the parent of one post. To improve the model accuracy and efficiency, we propose to materialize fully labeled instances considering the probability of each post being a candidate parent. The challenge is how to estimate the probability of a post being the parent of another post.

As person references give hints on parent-child relationship between posts, we propose to use the person resolution techniques to evaluate the likelihood of all possible candidate parents and only materialize the most likely candidates. In this way, we expect the materialized thread structures are more accurate, and help learn a more accurate threadCRF model. When there is no clear person resolution indication for some posts, we use three unsupervised rule-based baseline approaches: reply to the first post, reply to the last post, and reply to the post with the highest content similarity, referred as FIRST, LAST, and SIM, respectively.

The whole materialization process is shown in Algorithm 2. The complexity is $O(m^2L + 3^m)$, where $m$ is the number of posts in a thread and $L$ is the number of PR feature types. In the worst case, it will generate at most $3^m$ training instances.

**Encoding PR into ThreadCRF.** Person resolution not only can help generate fully labeled training data set to train the threadCRF model, it can also be incorporated as part of the threadCRF model. Recall that the threadCRF model includes a set of node and edge features. We encode person resolution as a node feature. Specifically, we define the PR feature value $PR(y_i = j)$ between post $i$ and its candidate parent $j$, where $i \geq 1$, $j \geq 0$, and $i > j$. We assign a weight for each PR feature type based on its priority. Assume there are $L$ types arranged in the descending order of priority. For the $k$th type, $1 \leq k \leq L$, its weight is assigned as $\exp(L - k)$. Suppose we are evaluating the PR feature value for the $i$th post to reply to the $j$th post, represented
Algorithm 2: Materialization with PR Evaluation

**Input**: One thread with a partially known reply structure.

**Output**: Multiple thread instances each with a complete reply structure.

1. for each post (starting from the third post) do
2.   if its parent post is unknown then
3.     List all the preceding posts as the candidate parents;
4.     Use the pseudo codes from Line 4 to Line 12 in Algorithm 1 to shrink the candidate set;
5.     if the candidate set includes the first, last, and the post with the highest content similarity then
6.       Further shrink the candidate set by only retaining the first, last, and the most similar post as the candidates;
7.   else
8.     List the known parent as the only candidate.
9.    end
10. Generate all the possible thread structures by picking one candidate for each post (starting from the third post).

as \( PR(y_i = j) \), we have

\[
PR(y_i = j) = \sum_{k=1}^{L} \exp(L - k) \cdot \delta_{ijk}, \tag{2.7}
\]

where,

\[
\delta_{ijk} = \begin{cases} 
1 & \text{if the } k\text{th type is matched for post } i \text{ and its parent } j; \\
0 & \text{Otherwise.} 
\end{cases} \tag{2.8}
\]
2.3 Evaluation

In this section, we evaluate the proposed system and approaches.

2.3.1 Thread Structure Learning Evaluation

In this subsection, we present the data sets, comparison methods, evaluation metrics, results, and analysis of experimental evaluation of the proposed thread structure learning methods.

Data Sets

We evaluate the proposed methods with two different health forum data sets.

Patients Forum Data Set. Although most of health forums only have partially labeled structures, we managed to find one forum, the patients forum on tumors of the parotid gland (http://patientsforum.com), that has fully labeled thread structures with each represented in a hierarchical tree view. We collected all 23842 posts in 2646 threads. In our experiments, we randomly removed a known reply relationship with probability 0.3, and get 5561 unknown reply relationships. Note that given any data set, our approach can always use part of them for training the model, and test it on the whole data set. Unless otherwise specified, we select 1105 threads with more labeled relationships to compose the training set, and use the proposed methods to predict the removed reply relationships in all 2646 threads. We compare with the original data for performance evaluation.

Epilepsy Forum Data Set. We also collected 9210 posts in 911 threads (topics) published on “Patient help patient” sub-forum in the previous mentioned epilepsy foundation discussion forum. In this forum, some posts explicitly reply to a preced-
Table 2.1: Comparison Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRST</td>
<td>Reply to the first post.</td>
</tr>
<tr>
<td>LAST</td>
<td>Reply to the last post.</td>
</tr>
<tr>
<td>SIM</td>
<td>Reply to the post with the highest content similarity.</td>
</tr>
<tr>
<td>PR</td>
<td>Reply to the post selected according to Algorithm 1.</td>
</tr>
<tr>
<td>MEP</td>
<td>Materialize all the possible training instances with equal probabilities, and then train and test threadCRF with the known structures as constraints.</td>
</tr>
<tr>
<td>MPR</td>
<td>Materialize the training set using Algorithm 2, and train and test threadCRF with the known structures as constraints.</td>
</tr>
<tr>
<td>EPR</td>
<td>Materialize the training set using Algorithm 2, and train and test threadCRF with the known structures as constraints plus the PR feature.</td>
</tr>
</tbody>
</table>

 ing post by quoting that post; while others have unknown reply relationships. As discussed earlier, the goal of this work is to leverage partially labeled thread structures to learn complete thread structures. We need to select a subset of threads that have more known reply relationships to train the threadCRF model. We have chosen 200 threads for the experiments. As it is very expensive to obtain an objective ground truth, we only manually labeled all the 468 unknown reply relationships in the selected 200 threads for evaluation.

Comparison Methods

We tested our proposed methods and compared them with existing methods. All the tested methods are divided into three categories: rule-based, CRF, and CRF + PR. Table 2.1 explains all those methods. Our proposed methods are marked in a bold font. For comparison methods, FIRST, LAST, and SIM have been used in [77]. MEP is the direct adaption of threadCRF for our proposed application scenario.
**Evaluation Metrics**

We define two categories of evaluation metrics, which are the same as those defined in [77] except that we only need to evaluate on the partially unknown structures. The first category is about the accuracy of individual parent labels or paths from the node to the root in the thread structure tree. The **accuracy of individual labels**, denoted as $\text{Acc}_{\text{edge}}$, is defined as the proportion of correct labels in the whole set of predicted labels. Let $U$ denote the set of posts with unknown parent labels, $\bar{y}$ denote the ground-truth label for $p_i \in U$, and $\hat{y}_i$ denote the predicted label for $p_i$. We define

$$\text{Acc}_{\text{edge}} = \frac{\sum_{p_i \in U} \delta[\bar{y}(i) = \hat{y}(i)]}{|U|}, \quad (2.9)$$

where $|U|$ is the size of set $U$. $\delta[\bar{y}(i) = \hat{y}(i)] = 1$ if the two labels are the same. Otherwise, it is zero.

We also define the **path accuracy**, denoted by $\text{Acc}_{\text{path}}$, as the proportion of correct paths from each node to the root in the thread structure tree.

$$\text{Acc}_{\text{path}} = \frac{\sum_{p_i \in U} \delta[\text{path}(i) = \hat{\text{path}}(i)]}{|U|}, \quad (2.10)$$

where $\text{path}(i)$ and $\hat{\text{path}}(i)$ are the set of nodes (posts) in the path from the node $i$ (post $p_i$) to the root node in the ground-truth path and the predicted path, respectively. $\delta[\text{path}(i) = \hat{\text{path}}(i)] = 1$ if the two paths are identical. Otherwise, it is zero. Note that the path-based metrics emphasize that correct prediction of the labels for those nodes with more descendants is more important.

In the second category, we define the **path-based precision and recall**, which are a relaxation of the accurate path matching as in Eq. 2.10. The precision is the proportion of the predicted paths that are part of the ground-truth paths in all the
predicted paths. The recall is the proportion of the ground-truth paths that are part of the predicted paths in all the ground-truth paths. They are mathematically defined as follows.

\[
P_{\text{path}} = \frac{\sum_{p_i \in U} \delta[\hat{\text{path}}(i) \subseteq \text{path}(i)]}{|U|}, \tag{2.11}
\]

\[
R_{\text{path}} = \frac{\sum_{p_i \in U} \delta[\text{path}(i) \subseteq \hat{\text{path}}(i)]}{|U|}, \tag{2.12}
\]

where \(\delta[\hat{\text{path}}(i) \subseteq \text{path}(i)] = 1\) if \(\hat{\text{path}}(i)\) is a subset of \(\text{path}(i)\). Otherwise, it is zero.

We also define \(F_{1\text{path}}\) as the harmonic mean of \(P_{\text{path}}\) and \(R_{\text{path}}\).

For each defined metric, there are two levels of evaluation: thread level and corpus level, which are used in [77]. In the thread level, these metrics are first measured for each thread, and then they are averaged through all the threads in the test set. It emphasizes the thread structure learning performance for each thread. In the corpus level, these metrics are directly evaluated for the whole test set without the thread-level evaluation and aggregation process.

**Results and Analysis**

In this section, we show and analyze the experimental results. Table 2.2 and Table 2.3 show the thread structure learning performance on the two data sets. We underline the numbers in each row if they are highest among the rule-based methods or among the CRF-based methods. The numbers in bold font represent the best performance among all the methods. Fig. 2.4 and Fig. 2.5 show the impact of training set size.

**Comparison among rule-based methods.** Table 2.2 and Table 2.3 show that, among the rule-based methods, \(PR\) achieves the best performance for most of the
Table 2.2: Performance Comparison on the Patients Forum Data Set

<table>
<thead>
<tr>
<th></th>
<th>Rule-based</th>
<th>CRF</th>
<th>CRF + PR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FIRST</td>
<td>LAST</td>
<td>SIM</td>
</tr>
<tr>
<td>Acc_edge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.427</td>
<td>0.362</td>
<td>0.376</td>
</tr>
<tr>
<td>corpus</td>
<td>0.360</td>
<td>0.363</td>
<td>0.336</td>
</tr>
<tr>
<td>Acc_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.427</td>
<td>0.119</td>
<td>0.276</td>
</tr>
<tr>
<td>corpus</td>
<td>0.360</td>
<td>0.086</td>
<td>0.217</td>
</tr>
<tr>
<td>P_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>1.000</td>
<td>0.119</td>
<td>0.535</td>
</tr>
<tr>
<td>corpus</td>
<td>1.000</td>
<td>0.086</td>
<td>0.461</td>
</tr>
<tr>
<td>R_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.427</td>
<td>1.000</td>
<td>0.520</td>
</tr>
<tr>
<td>corpus</td>
<td>0.360</td>
<td>1.000</td>
<td>0.457</td>
</tr>
<tr>
<td>F1_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.599</td>
<td>0.212</td>
<td>0.528</td>
</tr>
<tr>
<td>corpus</td>
<td>0.529</td>
<td>0.158</td>
<td>0.459</td>
</tr>
</tbody>
</table>

evaluation metrics. In Table 2.2, PR achieves the best performance for Acc\_edge, Acc\_path, and F1\_path. In Table 2.3, PR achieves the best performance for Acc\_edge and F1\_path. Note that FIRST has perfect P\_path since all the ground-truth paths have to contain the first post and itself, which are the only two posts in the predicted paths. In other words, all the predicted paths are part of the ground-truth paths, which leads to a perfect P\_path. The similar reason explains LAST’s R\_path performance. However, their F1 performances are worse than the PR method.

**The performance of CRF-based methods.** For the CRF-based methods, we evaluate the performance of MEP, MPR, and EPR. First, we found that in many cases, MEP is not as good as PR. In Table 2.2, MEP outperforms PR in Acc\_path, P\_path, and F1\_path, but achieves a slightly worse performance in Acc\_edge and R\_path. In
Table 2.3: Performance Comparison on the Epilepsy Forum Data Set

<table>
<thead>
<tr>
<th></th>
<th>Rule-based</th>
<th>CRF</th>
<th>CRF + PR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FIRST</td>
<td>LAST</td>
<td>SIM</td>
</tr>
<tr>
<td>Acc_edge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.528</td>
<td>0.383</td>
<td>0.405</td>
</tr>
<tr>
<td>corpus</td>
<td>0.444</td>
<td>0.429</td>
<td>0.361</td>
</tr>
<tr>
<td>Acc_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td><strong>0.528</strong></td>
<td>0.233</td>
<td>0.345</td>
</tr>
<tr>
<td>corpus</td>
<td>0.444</td>
<td>0.203</td>
<td>0.274</td>
</tr>
<tr>
<td>P_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td><strong>1.000</strong></td>
<td>0.233</td>
<td>0.675</td>
</tr>
<tr>
<td>corpus</td>
<td><strong>1.000</strong></td>
<td>0.203</td>
<td>0.647</td>
</tr>
<tr>
<td>R_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.528</td>
<td><strong>1.000</strong></td>
<td>0.608</td>
</tr>
<tr>
<td>corpus</td>
<td>0.444</td>
<td><strong>1.000</strong></td>
<td>0.521</td>
</tr>
<tr>
<td>F1_path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread</td>
<td>0.691</td>
<td>0.377</td>
<td>0.639</td>
</tr>
<tr>
<td>corpus</td>
<td>0.615</td>
<td>0.337</td>
<td>0.578</td>
</tr>
</tbody>
</table>

Table 2.3, MEP only outperforms PR in R_path. We analyzed the reason why MEP, which uses a machine learning model, is not satisfactory. MEP is unaware of person reference relationships in thread structures and assumes that all the possible reply structures are equally likely, which introduces lots of incorrect labels into the training data. The materialized fully labeled training set is thus not a good approximation of the ground truth, which leads to a less accurate trained model, and thus a lower quality.

When we combine person resolution with thread CRF, we can see that the performance is significantly improved. In terms of Acc_edge, Acc_path, and F1_path, MPR outperforms MEP, and EPR outperforms MPR for both data sets. For MPR, as we have used person resolution to select the more likely candidate parent posts during
the training set generation, the training set is more similar to the ground truth, which helps to learn a more accurate model. That explains why MPR outperforms MEP. By comparing MPR with EPR, we can clearly see that the threadCRF model with the additional PR feature has a better prediction performance. EPR consistently outperforms all the other methods in Acc\textsubscript{edge}, Acc\textsubscript{path}, and F1\textsubscript{path}.

The impact of training set size We also analyze the impact of the training set size on the prediction performance of two best methods: MPR and EPR. We randomly
selected a set of threads as the training set, and tested on all the threads in the data set. When the training set size is zero, it means that there is no training process and the feature weights are all set to 1.0. Fig. 2.4 and Fig. 2.5 show the thread-level performance. Note that the trend of the corpus-level performance is similar and not shown here. It shows that the increasing of training size can improve the performance at the beginning. In Fig. 2.4, the performance for both MPR and EPR has stopped improving after the size is larger than 500. In fact, as also observed in [77], with a small training set, threadCRF can already achieve an encouraging performance compared to a larger training set. Such findings show that, in order to bootstrap our training process and predict all the unknown thread structures, we only need a small set of threads that have a majority of labeled reply relationships. In Fig. 2.5, the performance for MPR does not continue improving with the training size increasing from 150 to 200, while the performance for EPR keeps increasing. It suggests that with more features, more training instances are needed.

2.3.2 Patient-Centered Mining Evaluation

In this subsection, we present the data set, ground truth, evaluation metrics, experimental results and analysis for the patient-centered mining system evaluation.

Data Set

We use multi-keyword queries to evaluate our patient-centered mining system. We use the publicly available data in the epilepsy foundation discussion forum, which is initiated and maintained by National Institute of Neurological Disorders and Stroke (NINDS). We collected 9210 posts included in 911 threads (topics) published on the “Patient help patient” sub-forum by Nov. 2011. In order to leverage real user queries without introducing bias, we follow the method used in [17] to randomly select queries.
First, we find all the thread titles in the forum that end with a question mark. Since such a title indicates that a user, the thread initiator, is looking for answers to a question, it naturally represents as a user query. We then extract keywords from these thread titles. Instead of using a stop word list to filter out unimportant words, we choose MetaMap tool to extract phrases as the query keywords. The reason is that we want to identify each medical phrase containing multiple words and treat it as a unit in query processing. We randomly chose ten such thread titles with each corresponding to one query. We only tested ten queries because it is extremely labor-intensive to generate the ground truth for each query, especially since some queries may involve a large number of threads, which may include an enormous number of posts. More evaluation of the patient-centered mining system will be introduced in Chapter 4.

**Ground Truth**

We manually find the ground truth of relevant results for each query, based on the analyzed user expectation as discussed in Chapter 1. We assume AND semantics among all the keywords in a query. To generate the ground truth for a query, we first define a relevant thread as a thread that contains all the query keywords. Consider the intensive human labor, we randomly choose 30 relevant threads for manual checking if a query involves more than 30 relevant threads. Since a patient is a semantic unit, we find the relevant patients whose associated information contains all the query keywords from the relevant threads. Then we consider the posts that are associated with such patients and contain at least one query keyword in the associated information as ground truth.
Table 2.4: Ten Randomly Chosen Queries

<table>
<thead>
<tr>
<th>Query questions (Keywords are underlined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

We compare the ground truth with post-based search, thread-based search, and our approach, referred as patient-based search. Post-based search returns all the posts each containing all the query keywords. Thread-based search returns all the posts each containing at least one query keyword in a relevant thread. Our patient-based search returns all the posts each containing at least one query keyword associated with a relevant patient. Note that our approach shares the same intuition as the ground truth, but automatically identifies patients and automatically associate information to each patient. The quality of these automated processes has been evaluated.
Evaluation Metrics

We use standard evaluation metrics in information retrieval: precision ($P$), recall ($R$), and f-measure ($F1$). Precision is the ratio of the number of correctly returned posts to the total number of returned posts. The recall is the ratio of the number of correctly returned posts to the total number of posts that should be returned according to the ground truth. f-measure is defined as the harmonic mean of precision and recall:

$$F1 = \frac{2\times P \times R}{P + R}.$$ 

Experimental Results and Analysis

The experimental results are shown in Table 2.5. It shows that post-based approach has almost perfect precision as in most cases keywords in the same post refer to the same patient and have close relationship, but it has very low recall. On the other hand, thread-based search achieves perfect recall since we do not consider the relationships of keywords in different threads, but it has a very low precision. In contrast, our patient-based search has good precision and recall in general, and achieves a much higher f-measure than the other two approaches.

We also analyzed the major reasons that affect our system performance. First, the performance of the current NLP tools, especially the co-reference resolution tool, is not perfect. Second, some forum acronyms cannot be recognized, like “my DD” cannot be identified as “my daughter”. Third, some patients cannot be identified by our system due to informal language used in a forum and the limited context. Some acronym identification [39], data-driven approaches for understanding informal language [83], and context-aware text mining techniques [50], can be used to improve the current system.
**Table 2.5:** Evaluation for Ten Randomly Chosen Queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Post-based</th>
<th></th>
<th></th>
<th>Thread-based</th>
<th></th>
<th></th>
<th>Patient-based</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>1</td>
<td>0.978</td>
<td>0.379</td>
<td>0.547</td>
<td>0.509</td>
<td>1.0</td>
<td>0.674</td>
<td>0.764</td>
<td>0.698</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>0.973</td>
<td>0.379</td>
<td>0.545</td>
<td>0.477</td>
<td>1.0</td>
<td>0.646</td>
<td>0.702</td>
<td>0.695</td>
<td>0.698</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>0.059</td>
<td>0.111</td>
<td>0.378</td>
<td>1.0</td>
<td>0.548</td>
<td>0.923</td>
<td>0.706</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>0.333</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.667</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>0.286</td>
<td>0.444</td>
<td>0.636</td>
<td>1.0</td>
<td>0.778</td>
<td>1.0</td>
<td>0.714</td>
<td>0.833</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>0.5</td>
<td>0.667</td>
<td>0.5</td>
<td>1.0</td>
<td>0.667</td>
<td>1.0</td>
<td>0.5</td>
<td>0.667</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>0.063</td>
<td>0.118</td>
<td>0.087</td>
<td>1.0</td>
<td>0.16</td>
<td>0.8</td>
<td>0.75</td>
<td>0.774</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
<td>0.429</td>
<td>0.6</td>
<td>0.28</td>
<td>1.0</td>
<td>0.438</td>
<td>1.0</td>
<td>0.714</td>
<td>0.833</td>
</tr>
<tr>
<td>9</td>
<td>1.0</td>
<td>0.534</td>
<td>0.696</td>
<td>0.349</td>
<td>1.0</td>
<td>0.518</td>
<td>0.716</td>
<td>0.658</td>
<td>0.686</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>0.3</td>
<td>0.462</td>
<td>0.455</td>
<td>1.0</td>
<td>0.625</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Overall</td>
<td>0.995</td>
<td><strong>0.326</strong></td>
<td><strong>0.491</strong></td>
<td><strong>0.417</strong></td>
<td>1.0</td>
<td><strong>0.589</strong></td>
<td><strong>0.851</strong></td>
<td><strong>0.674</strong></td>
<td><strong>0.752</strong></td>
</tr>
</tbody>
</table>
In this chapter, I introduced patient-centered mining for effective information discovery in health forums. A patient-centered mining system is first presented, which includes four major components: \textit{PersonIdentification}, \textit{PersonResolution}, \textit{PatientIdentification}, and \textit{HealthInformationAssociation}. As the thread structures, the reply relationships between posts, are very important for patient-centered mining, but most online forums only have partially labeled structures. A unified framework is proposed to learn the complete thread structures based on a statistical machine learning model, which leverage the existing partially known structures and the abundant person reference relationships in health forums. The effectiveness of the proposed approaches for thread structure learning has been verified on two health forum data sets. The effectiveness of the patient-centered mining for effective information discovery has been verified with the proposed keyword query experiments.
Chapter 3

PATIENT EXPERIENCE MINING

3.1 Introduction

The existing work on experience mining classifies a sentence as experience-containing sentence or not experience-containing sentence independently [63, 43], and does not consider the context of a sentence. However, the context information can be very important for accurately classifying a sentence. First, an individual sentence can be only part of a complete experience description, and it may have nothing to do with experience by itself. The extensive use of informal language, like incomplete sentences, make it even more difficult to classify each sentence separately. Consider this short sentence “ER+/PR+ HER2-.” It is a breast cancer type, and has nothing to do with patient experience. Now let us check its local context: “A little about me. Found knot in July … Found 0.09mm invasive area. No lymph node involvement. ER+/PR+ HER2- . Onco score of 20 1/2. Now on tamoxifen. A survivor.” With its context sentences, we can recognize that breast cancer type as part of the patient’s personal experience. Second, the same sentences can be either part of experience or not part of experience given different global context, e.g., different post authors. Consider a post with two sentences, “Keppra could cause birth defects. It is really not a safe drug.” Such sentences are more likely to be part of patient experience if they are from a post written by a patient who reported the drug usage experience in an earlier post other than a post written by a doctor who rarely shares any personal experience.

We propose to exploit the sentence context information for experience mining.
Suppose we have a set of threads from an online health forum. Each thread consists of a sequence of posts. Each post consists of a sequence of sentences, and is written by a forum user. We extract a set of sentences $S = \{s_1, s_2, ..., s_l\}$ from the threads with the Stanford NLP toolkit [56]. Each sentence $s_i$ has its author context $u_i$, e.g., whether the author is a doctor, post context $p_i$, e.g., the neighboring sentences in the same post, and thread context $t_i$, e.g., the post position in the thread.

We model the problem as a classification task. Let $Y = \{y_1, y_2, ..., y_l\}$ be the corresponding labels, where $y_i \in \{-1, 1\}$, $y_i = 1$ indicates that $s_i$ is a sentence containing patient experience, and $y_i = -1$ indicates that $s_i$ is not a sentence containing patient experience. Each sentence $s_i \in S$ has its context information $\langle u_i, p_i, t_i \rangle$. We aim to learn a classifier from labeled sentences with their content, context and labels, which can predict the labels for the unlabeled sentences.

3.2 Context Information for Patient Experience Mining

In this section, we introduce two types of context information: global context and local context. The global context is the context information shared by the whole post, while the local context is the context information within a post. Later we propose a Context-Aware expeRience Extraction (CARE) framework to incorporate the two types of context information.

The global context of a sentence is the context information of the post containing that sentence, which is used to differentiate the posts that are more likely to contain patient experience from the other posts. It can be extracted from the post author context or the thread context of a post.

As different post authors may have different preferences about sharing personal experience, we can take the author context or profile into consideration when classifying a sentence posted by that author. For example, a patient tends to share their
personal experience, while a doctor rarely does that. We can thus use the author type, whether the author is a doctor or not as an author context feature. More author context features, such as the number of replies posted by the author and the number of badges the author has won, can also be considered.

The thread context of a post can also help extracting the patient experience sentences. For example, the post position, in particular, whether it is the first post in a thread, can be used as a thread context feature. The first post tends to share more personal experience, while the following posts tend to contain some suggestions and comments from the other forum users. We can also consider whether a post is from the thread initiator, i.e., the author of the first post. If a post is from the thread initiator, it is more likely to share some personal experience.

The local context refers to the context information within a post. Consider the following example from a forum post. We use Stanford NLP for sentence splitting, and each sentence is preceded by its sentence sequence number.

“(1) I take Sotalol. (2) It is ok in preventing ventricular arrhythmias. (3) especially on high dosage. (4) After 18 months of assumption my ICD recorded only a few Vtachs, while when I was on Coreg I had at least one Vtach a week. (5) Your bpm and bp seem just ok. (6) You want them lower? (7) You can... ”

In this example, the first 4 sentences are patient experience sentences, while sentence (5)-(7) are not. We notice that sentence (1) and (4) can be identified as patient experience sentences much easier than sentence (2) and (3). The reason is that both sentence (1) and (4) contain some distinguishable features. For example, as shown in the bold font, sentence (1) uses the pronoun “I” as the subject and sentence (4) further uses the verbs in their past tense. Such sentence-level features are useful to identify them as patient experience. On the other hand, it is difficult to identify (2) and (3) when we only consider themselves; however, we can identify them as patient experience.
experience with their context sentences (1) and (4).

Aforementioned analysis suggests that the label of a sentence tends to be consistent with the labels of its local context sentences. Such local context may be leveraged to improve the patient experience sentence extraction performance. More specifically, we make use of local context based on the following two observations.

• Observation 1: The adjacent sentences tend to have the same label.

• Observation 2: The sentences in the same post tend to have the same label.

The first observation follows the facts that the experience or non-experience sentences are usually contiguous. In many cases, a sequence of patient experience sentences follow or are followed by a sequence of non-experience sentences. For example, a forum user can make some comments or suggestions in the first half of a post as a reply to a preceding post, and then describe some of her own experience. Or, she can describe her own experience first, and then give some suggestions, as shown in the above example.

The second observation follows the facts that the label of an individual sentence usually depends on the labels of the other sentences in the same post. If most of the other sentences in that post are patient experience sentences, then it is more likely that the individual sentence is also a patient experience sentence. Otherwise, if none of the other sentences are patient experience sentences, then it is less likely one individual sentence is a patient experience sentence.

We have verified the above two observations based on statistical analysis. We crawled the data from the MedHelp Heart Disease forum, and randomly selected a 1

---

1We may make other observations about local context; for example, sentences within a distance or containing co-referent entities tend to have the same label, and we would leave it as our future work.
subset of threads as our data set. One undergraduate student and one graduate student both in Computer Science major were hired to manually label all the sentences in 264 posts from 85 threads. Among the 264 posts, 105 posts do not contain any patient experience sentences, which account for 41.6% of all the posts. The total number of sentences we get from the 264 posts is 2007 by using Stanford NLP for sentence splitting. By removing the sentences with less than 3 tokens, e.g., a sentence with only a single punctuation mark, and the question sentences, which are considered as non-experience, we get 1481 sentences. As shown in Table 4.1, two random sentences have the same label with the probability 0.5. Two adjacent sentences (from the same post) have the same label with the probability 0.865. Two sentences from the same post have the same label with probability 0.815. The statistics show that the adjacent sentences or the sentences from the same post indeed tend to share the same label. These observations are used to model local context in the problem of patient experience extraction.

### 3.3 Context-Aware Experience Extraction

We use SVM as the basic model for the proposed framework CARE to classify each sentence as containing patient experience or not containing patient experience due to
its superior performance in many real-world applications [16]. Let \( f = \{f_1, f_2, ..., f_{k_1}\} \) be the set of \( k_1 \) sentence-level features extracted from the content of sentences. Let \( Z = \{z_1, z_2, ..., z_l\} \in \mathbb{R}^{k_1 \times l} \) be the feature matrix representation for \( l \) labeled sentences \( \{s_1, s_2, ..., s_l\} \). Each feature vector \( z_i, 1 \leq i \leq l \), contains all the \( k_1 \) feature values extracted from sentence \( s_i \). We also have \( Y = \{y_1, y_2, ..., y_l\} \) as the corresponding labels, where \( y_i = 1 \) indicates sentence \( s_i \) is a patient experience sentence, and \( y_i = -1 \) if it is not. Given \( Z \) and \( Y \), we learn a standard soft-margin support vector machine by solving the following optimization problem.

\[
\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\
\text{s.t.} \quad y_i (w^T z_i + b) \geq 1 - \xi_i, \\
\quad \xi_i \geq 0, i = 1, \ldots, l, 
\]

where \( w \) is the feature weight vector, and \( b \) is the bias. \( \xi_i \) is the soft-margin slack variable for \( z_i \), which allows the noise in the training sample. \( C \) controls the total error from training samples.

The global context, such as author context and thread context, can be considered as features of each sentence, which suggests that we can exploit global context via feature engineering. Let \( g = \{g_1, g_2, ..., g_{k_2}\} \) be the set of \( k_2 \) features extracted from global context. We expand the sentence-level feature set \( f \) with the context feature set \( g \), and let \( F = f \cup g \) be the feature set of \( k = k_1 + k_2 \) features for each individual sentence. Let \( X = \{x_1, x_2, ..., x_l\} \in \mathbb{R}^{k \times l} \) be the new feature matrix representation for \( l \) labeled sentences \( \{s_1, s_2, ..., s_l\} \) with the feature set \( F \). Each feature vector, \( x_i, 1 \leq i \leq l \), contains \( k_1 \) sentence-level features and \( k_2 \) features from its global context.

Different from the global context, it is difficult to model local context via feature
engineering, hence we model the local context as labeling constraints between sentences within a post. Our previous analysis suggests that two adjacent sentences or two sentences from the same post tend to share the same label. We define a matrix $A^N \in \mathbb{R}^{l \times l}$ to encode the adjacent-sentence relationship where $A^N_{ij} = 1$ if sentences $s_i$ and $s_j$ are adjacent sentences and $A^N_{ij} = 0$ otherwise. Similarly, we can introduce a matrix $A^P \in \mathbb{R}^{l \times l}$ to encode the same-post relationship where $A^P_{ij} = 1$ if sentences $s_i$ and $s_j$ are from the same post and $A^P_{ij} = 0$ otherwise. The following derivations are based on the sentence-sentence label consistency matrix $A \in \mathbb{R}^{l \times l}$. We can obtain $A$ from either $A^N$, $A^P$ or their combination as $A = A^N + \lambda A^P$, where $\lambda$ controls the weight of different local context in the model. In this work, we focus on studying the effects of contextual information on patient experience extraction performance, but not ways to combine them. Therefore, we simply combine them with equal weight $\lambda = 1$ to construct a sentence-sentence label consistency matrix.

To model local context for patient experience extraction, the basic idea is to make two sentences likely to share the same label if they are two adjacent sentences or from the same post. The basic idea can be mathematically formulated as solving the following minimization problem:

$$
\min_w \frac{1}{2} \sum_{i,j} A_{ij}(w^T x_i - w^T x_j)^2 = w^T XLX^T w, \quad (3.2)
$$

where $X \in \mathbb{R}^{k \times l}$ is the matrix representation of sentences in the training set. $L = D - A$ is a Laplacian matrix where $D$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{l} A_{ij}$.

With model components for global and local context, the proposed patient experience extraction framework CARE is to solve the following optimization problem:

$$
\min_{w, b, \xi} \frac{1}{2} w^T w + C_e \sum_{i=1}^{l} \xi_i + C_r w^T XLX^T w \\
\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \\
\xi_i \geq 0, i = 1, \ldots, l, \quad (3.3)
$$
where $C_r$ is the parameter to control the contribution from local context.

We can easily collect a large amount of unlabeled sentences from online health forums. The proposed framework CARE can be extended to incorporate unlabeled sentences. Let $S^U$ be a set of $m$ unlabeled sentences. We use $X^U \in \mathbb{R}^{k \times m}$ to denote the matrix representation of $S^U$ and $X' = [X, X^U] \in \mathbb{R}^{k \times (l+m)}$ to denote labeled and unlabeled sentences. Similarly, we extract features to model global context and construct the sentence-sentence label consistency matrix $A' \in \mathbb{R}^{(l+m) \times (l+m)}$ to model local context. The formulation extending CARE to incorporate unlabeled data is as below:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C_e \sum_{i=1}^{l} \xi_i + C_r w^T X' L' (X')^T w$$

s.t.  \hspace{1cm} y_i(w^T x_i + b) \geq 1 - \xi_i,

$$\xi_i \geq 0, i = 1, \ldots, l,$$

where $L'$ is the Laplacian matrix built based on $X'$ for both labeled and unlabeled data [91]. The optimization problems in Eqs. (3.3) and (3.4) can be solved by the primal Laplacian SVM solver [10, 57].

3.4 Experiments and Analysis

In this section, we conduct experiments to answer the following two questions - (1) can the proposed framework CARE help patient experience extraction? and (2) how does contextual information affect the performance of CARE? We begin by introducing experimental settings.

3.4.1 Experimental Settings

We use the same data set as described in Section 3.2. In that data set, we have 2007 labeled sentences in 264 posts from 85 threads crawled from the MedHelp forum.
After removing the sentences with less than 3 tokens and the question sentences, we got 1481 sentences as our experiment data set.

We randomly divide all the posts into two subsets of equal size $A$ and $B$. All the sentences from $A$ are used for training, and the others from $B$ are used for testing. We always fix $B$ as the testing set; while choose $\alpha\%$ of $A$ as labeled data and the remaining $1 - \alpha\%$ as unlabeled data [54]. We vary $\alpha$ as $\{10, 25, 50, 100\}$ in this work. We draw 10 random splits and report the average classification accuracy. Here the accuracy is defined as the ratio of the number of correctly classified sentences to the total number of sentences in the testing set.

We use two global context features: $CtxA$ and $CtxT$, and one regularization term for local context: $Reg$. $CtxA$ is extracted from the author context: whether the post author is a doctor or not. There are many doctors who help answer questions and give suggestions to patients in the MedHelp forums. Their profile webpages usually indicate that they are doctors. We crawled all the profile webpages for those forum users who are a post author in our data set, and extract such context feature. $CtxT$ is extracted from the thread context of a post: whether the post is the first post in the thread. We use the regularization defined in Section 3.3.

### 3.4.2 Baselines

We compare our work with representative baselines, which also user SVM as basic classifier but with different sets of sentence-level features.

**Bag of Words, Bigram, POS tags**

N-gram and POS tag features are commonly used for text classification. [43] has used bag of words (BOW, also known as unigram), Bigram, and BOW + POS features to distinguish personal experience from hearsay. We use their approaches as
baselines. We also use Weka [23] to generate a default n-gram combination of 1000 most important unigram, bigram, or trigram features for each sentence. We denote such feature set as $\text{N-gram(Weka)}$.

**Linguistic features**

Most of the existing work for the experience extraction and classification task uses a set of linguistic features extracted from each sentence [63, 60]. We follow them and use three sets of linguistic features, totally 21 features, as our sentence-level features. We denote them as $\text{SentLing}$.

- **Pronouns**: As shown in [60], the presence of pronouns is very useful in identifying personal experience revealing sentences. We follow the approaches in [60] to design 8 features by extracting pronouns in 8 pronoun categories.

- **Tense**: As shown in [63, 60], the tense of a sentence is very helpful in identifying personal experience. Most of the experience descriptions are using the past tense or present tense. We follow [63] to design 6 tense features.

- **Modality**: As shown in the same existing work as above, the modality expresses the possibility of some activities or events, and thus it can be used to differentiate experience from non-experience. We follow [63] to design 7 modality features.

The parameters of SVM for all baseline methods are determined by cross validation. For the proposed framework CARE, we use the linear kernel, and set $C_e = 0.00001$ and $C_r = 0.05$. We will give more details about parameter analysis for CARE in the following subsections.
3.4.3 Performance Comparison

The comparison results are shown in Table 3.2. Note that we use features in SentLing as the feature set for the content of sentences for the proposed framework CARE. We make the following observations:

- BOW and BOW + POS always outperform Bigram. The feature sets of Bigram are much larger than BOW and BOW + POS, which will degrade the performance of SVM because of curse of dimensionality.

- SentLing obtains better performance than the other baselines, which suggests the importance of linguistic features in patient experience extraction.

- The proposed framework CARE can significantly improve the performance of patient experience extraction. For example, compared with the best performance of baseline methods, CARE improves the performance by 6.0% with 10% labeled data and 4.3% with 100% labeled data. The major reason is that the proposed framework captures both global and local context. In the following subsection, we will investigate the impact of context information on CARE.

In summary, with the help from context information, the proposed framework CARE can significantly improve the performance of patient experience extraction.

3.4.4 Impact of Context Information

The proposed framework exploits global and local context to improve the performance of patient experience extraction. In this subsection, we investigate the effects of context information on the performance of CARE by systematically defining its variants:
• CARE-B: it only uses the basic classifier SVM with the SentLing features without any context information;

• CARE-A: it combines SentLing features and only CtxA features from global context;

• CARE-G: it combines SentLing features, and both CtxA and CtxT features from global context;

Parameters in these variants are determined via cross-validation and the comparison results are shown in Figure 3.1. First, CARE-A outperforms CARE-B, which supports the importance of CtxA features from global context especially when more labeled data is available. Second, CARE-G improves the performance by combining CtxA and CtxT features from global context, which suggests that CtxA and CtxT contain complementary information. Third, CARE can further improve the performance by incorporating local context into CARE-G especially when less labeled data is available. For example, the accuracy increases from 0.797 to 0.823 with 10% labeled training data. These observations suggest that both global and local context are useful for the performance improvement for patient experience extraction.

3.4.5 Parameter Analysis

The proposed framework CARE has one important parameter $C_r$, which controls the contribution from local context. Fig. 3.2 shows how the performance of CARE varies with the change of the parameter $C_r$. When the labeled data is small such as 10% of labeled data, with the increase of $C_r$ from 0 to 0.05, the performance increases dramatically. With the further increase of $C_r$ from 0.05 to 0.1, the performance slightly decreases but is still much better than $C_r = 0$. This observation has its practical significance because it is time and effort consuming to obtain labeled data.
Table 3.2: Performance Comparison

<table>
<thead>
<tr>
<th>Features</th>
<th>Labeled Data</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-level</td>
<td>BOW</td>
<td>0.721</td>
<td>0.758</td>
<td>0.77</td>
<td>0.785</td>
</tr>
<tr>
<td></td>
<td>Bigram</td>
<td>0.531</td>
<td>0.642</td>
<td>0.703</td>
<td>0.740</td>
</tr>
<tr>
<td></td>
<td>BOW + POS</td>
<td>0.728</td>
<td>0.77</td>
<td>0.779</td>
<td>0.795</td>
</tr>
<tr>
<td>N-gram (Weka)</td>
<td>0.719</td>
<td>0.755</td>
<td>0.758</td>
<td>0.766</td>
<td></td>
</tr>
<tr>
<td>SentLing</td>
<td>0.777</td>
<td>0.797</td>
<td>0.809</td>
<td>0.819</td>
<td></td>
</tr>
<tr>
<td>CARE</td>
<td>0.823</td>
<td>0.842</td>
<td>0.848</td>
<td>0.854</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.1: The Impact of Context Information

while we can easily get a large amount of unlabeled data. With more labeled data, the performance is relatively stable. In some cases such as 100% of labeled data, we cannot even observe any improvement with the regularization constraints. It suggests that the regularization constraints for local context work best when there is little labeled data but a large amount of unlabeled data.
In this chapter, I presented context-aware experience extraction from online health forums. The extraction of patient experience is modeled as a classification problem: classifying each sentence in a forum post as containing patient experience or not containing patient experience. The sentence context information is exploited for such experience extraction task. The context information is first classified into two types: global context and local context, and then incorporated into a unified context-aware experience extraction framework CARE. The experimental results show that the global context can significantly improve the experience extraction accuracy, while the local context can also improve the performance when less labeled data is available.
4.1 Introduction

Adverse Drug Reactions (ADRs) have become a serious health problem and even a leading cause of death in the United States. The Institute of Medicine reported in January of 2000 that an estimated 7,000 deaths occur due to ADRs [30]. A systematic overview conducted in 2007 shows that 5.3% of hospital admissions were associated with ADRs [31]. It is estimated that more than two million patients in the United States are affected each year by ADRs, making ADR the fourth leading cause of death in the United States [35, 12, 82].

Pre-marketing review and post-marketing surveillance have been used to identify ADRs. Pre-marketing review is required before any drugs are approved for marketing, which mainly relies on clinical trials with a selected set of patients to identify the potential drug risks. Due to the limited sample size and duration of clinical trials, pre-marketing review is insufficient for identifying all the potential ADRs for various patients. Post-marketing surveillance mainly relies on patient drug usage surveys or voluntary and spontaneous report systems [2]. However, according to [24], the median under-reporting rate across 37 studies using a wide variety of post-marketing surveillance methods from 12 countries is 94%.

On the other hand, online health forums may provide valuable information in a large scale and in a timely fashion through the active participation of patients and caregivers. Online health forums, such as MedHelp, WebMD, PatientsLikeMe, Healthboards, and etc., are a popular platform for people to share their personal
experience, participate in discussions, express their feelings, and to support each other in healthcare. The user population of such forums are rapidly growing. For instance, MedHelp currently has 13 million active monthly users. With highly valuable patient-contributed information that is ever increasing volume, health forums provide the potential to discover knowledge about various disease treatments, their effects and adverse reactions, and so on.

In light of this, there is growing interest in leveraging online health forums to discover ADRs. The basis of ADR discovery is through co-occurrence analysis. The rationale behind such analysis is that if a drug can cause an adverse reaction, then they should be mentioned together frequently. Conversely, if a drug and an adverse reaction co-occur frequently, then the adverse reaction is likely to be related to the drug. Different studies consider different information units for co-occurrence analysis. For instance, in [11], the drug and the adverse reaction that co-occur within a window of 20 tokens in the same post are considered as potentially related. More recently, in [43, 44, 45], the drug and the adverse reaction that co-occur in the same sentence are considered as potentially related. In [85, 87], the drug and the adverse reaction that co-occur in the same thread are used for ADR discovery. Other machine learning techniques are used for ADR discovery, which are either based on the analysis on a sentence [61], a post [67, 88], or a thread information unit [81].

However, all of such information units are based on the syntax of documents, instead of the semantics. Specifically, such co-occurrences of the drug and the adverse reaction may not refer to the same patient. Intuitively, drugs taken by one person can not cause adverse effects experienced by another person. Thus ADR discovery should be based on the occurrence of a drug and an ADR at the same patient.

Semantic unit, patient, is typically not aligned with syntactic unit. The drug and its corresponding adverse reactions do not have to be mentioned in the same
sentence, post, or even thread. For example, one patient mentioned that she was on aspirin in one post. In another post, she mentioned that she always felt dizzy and frequently experienced spinning sensation. It is possible that dizzy or spinning sensation was caused by the drug aspirin she was taking despite the fact that their occurrences are in different posts. On the other hand, an adverse reaction cannot be caused by a drug if it was taken by a different patient even if they are mentioned in the same document unit. Consider a thread in which many patients shared their own experience. One patient mentioned that she had flushed skin, while another patient shared some different experience and also indicated that he was taking aspirin. If the first patient did not mention that she was on aspirin, we cannot make any inference that aspirin may cause flushed skin just because they co-occur in the same thread. Therefore, we need to identify all the drug and the adverse reaction pairs that are associated with the same patient for potential ADR discovery.

Another challenge of mining health forum for ADRs is varying degree of information quality. Some information includes patient-reported personal health experience, such as “Took Keppra for a week. I felt very nauseous”, which provides new ADR evidences. On the other hand, there is also a lot of hearsay on the forum, such as “Your daughter should not take Keppra as it is known to be unsafe for children.” Such unverified hearsay, if used for ADR discovery, can be misleading. To discover trustable ADR knowledge from online health forums, it is critical for us to extract patient-reported personal health experience and differentiate it from hearsay or advices. In [43], report source classification is proposed to differentiate patient experience from hearsay. In [44, 45], an evaluation based on 400 sentences manually labeled as patient experience or hearsay shows that differentiating them can significantly improve the adverse drug event extraction performance.

In this work, we proposed to use patient-centered and experience-aware mining for
effective ADR discovery. First, we have developed a patient-centered mining system, which can identify the mentioned patients in health forums, aggregate, and associate the health information with the corresponding patients. Second, we have developed an automatic patient experience mining system to distinguish patient experience from the other forum information. We integrated them to build a patient-centered and experience-aware health information database, which is used for effective ADR discovery. Fig. 1.1 has already shown the overall framework. We verified its effectiveness via experimental evaluation with an official ADR knowledge base as well as human-annotated ground truth [51].

4.2 Adverse Drug Reaction Discovery in Online Health Forums

In this section, we briefly introduce the background, including some underlying techniques used in existing work for ADR discovery in online health forums.

4.2.1 Co-occurrence Analysis

Co-occurrence analysis has been extensively used to discover ADRs in online health forums. For co-occurrence analysis, it assumes that if a drug and an adverse reaction co-occur frequently in the same information unit, then the drug is considered likely to cause the adverse reaction. The information unit can be a sentence, a post, a thread, or a patient health information unit. In this work, we consider the co-occurrence in the same patient health information unit.

Support and confidence are two common measures used for co-occurrence analysis [85, 87]. In general, the higher the support or confidence value, the more likely the drug and the adverse reaction are related. Therefore, we can identify the most likely ADRs based on the top-ranked support or confidence values. Let total\text{Count} denote the total number of information units in the data set, such as the total number
of sentences, posts, threads, or patients. Let \( \text{count}(D \cup R) \) denote the number of information units that contain both the drug \( D \) and the adverse reaction \( R \). The support is defined as follows.

\[
\text{support}(D \cup R) = \frac{\text{count}(D \cup R)}{\text{totalCount}} \tag{4.1}
\]

Let \( \text{count}(D) \) be the number of information units that contain the drug \( D \). The confidence is defined as follows.

\[
\text{confidence}(D \Rightarrow R) = \frac{\text{support}(D \cup R)}{\text{support}(D)} = \frac{\text{count}(D \cup R)}{\text{count}(D)} \tag{4.2}
\]

4.2.2 Evaluation Methods

The evaluation of the ADR discovery can be based on the comparison between the identified ADRs in health forums and the ADRs listed in an official knowledge base or a human-annotated ground truth. An official ADR knowledge base, such as FAERS and SIDER [32], is often used as ground truth for performance evaluation. FAERS refers to the US Food and Drug Administration (FDA) Adverse Event Reporting System, which is a database that contains information on adverse event and medication error reports submitted to the FDA. SIDER contains information on marketed medicines and their recorded adverse drug reactions. It also includes the drug indications which show what the drug is used to treat. We use both SIDER and a human-annotated ground truth to evaluate our proposed methods, and will discuss their advantages and disadvantages in the evaluation section.

4.2.3 Term Mapping

There is a gap between the consumer terms used in health forums and the professional terms used in the official ADR database like SIDER. The consumer health vocabulary (CHV) [89] has been used in the existing work to expand a drug or ADR
professional term in the knowledge base with all the equivalent expressions in the consumer vocabulary. In this work, we also use CHV to map the professional terms in SIDER to the consumer terms used in health forums. Each professional term in SIDER or consumer term in CHV has a concept unique identifier (CUI) in the Unified Medical Language System (UMLS) [41]. All the terms with the same CUI will be considered equivalent during our ADR discovery process. For example, the CUI of “diarrhea” is “C0011991”, which corresponds to 12 terms in CHV, including “diarrhea”, “loose bowel movement”, and etc. All these expressions will be considered as the same ADR.

4.3 Experimental Evaluation

In this section, we show the experimental evaluation by comparing our proposed methods with the existing methods.

4.3.1 Data Set

We collected 120275 posts with 998637 sentences included in 34065 threads publicly available in the MedHelp heart disease forum by Sep. 2014. 33826 forum users participated in the forum discussion, who either initiated a thread or replied to others as a post author. The detailed forum statistics and distributions are shown in Table 4.1. Notice that in the table, a forum user, who participated in the forum discussion by posting, is also a post author, and vice versa.
Table 4.1: MedHelp Heart Disease Forum Data Set Statistics

<table>
<thead>
<tr>
<th>Range</th>
<th>≤5</th>
<th>(5,10]</th>
<th>(10,20]</th>
<th>(20,30]</th>
<th>(30,50]</th>
<th>(50,100]</th>
<th>&gt;100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution of # of posts in a thread</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of threads</td>
<td>29128</td>
<td>3259</td>
<td>1181</td>
<td>350</td>
<td>114</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>Distribution of # of authors in a thread</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of threads</td>
<td>31844</td>
<td>1628</td>
<td>490</td>
<td>64</td>
<td>25</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Distribution of # of posts by an author</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of authors</td>
<td>31712</td>
<td>1099</td>
<td>525</td>
<td>169</td>
<td>134</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td>Distribution of # of threads a forum user participated in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of forum users</td>
<td>32541</td>
<td>628</td>
<td>354</td>
<td>94</td>
<td>80</td>
<td>69</td>
<td>60</td>
</tr>
<tr>
<td>Distribution of # of threads initiated by a forum user</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of forum users</td>
<td>24081</td>
<td>279</td>
<td>110</td>
<td>27</td>
<td>16</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Distribution of # of posts a forum user posted as a reply</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of forum users</td>
<td>16181</td>
<td>674</td>
<td>337</td>
<td>116</td>
<td>107</td>
<td>85</td>
<td>81</td>
</tr>
</tbody>
</table>

We use the whole data set for our large-scale evaluation with SIDER, and sample a small set for the evaluation with human annotation.

4.3.2 Ground Truth

We use two types of ground truth to evaluate the performance of different methods. The first type of ground truth is the official ADR knowledge base: SIDER, which allows us to evaluate the proposed methods in a large-scale fashion and with negligible human effort. The second type of ground truth is human annotation. Although the second one is labor-intensive, it is still extensively used in the existing work, as it provides some advantages that the first one lacks, which will be discussed later.
4.3.3 Comparison Methods

Table 4.2 shows the notation for three baseline methods: *Sent* [44], *Post* [67], and *Thread* [87], and four variants of our proposed method: *PI*, *PA*, *PIE*, and *PAE*. As discussed earlier, the three baseline methods or their variants have been widely used in the existing work. We use our proposed experience learning approach to classify each sentence in our data set as describing patient experience or not. If a drug or ADR is contained in an experience description sentence, we consider it as experienced.

**Table 4.2: Notation of the Comparison Methods**

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent</td>
<td>Co-occurrence analysis with a sentence as the information unit.</td>
</tr>
<tr>
<td>Post</td>
<td>Co-occurrence analysis with a post as the information unit.</td>
</tr>
<tr>
<td>Thread</td>
<td>Co-occurrence analysis with a thread as the information unit.</td>
</tr>
<tr>
<td>PI</td>
<td>The Patient-centered and In-thread method: co-occurrence analysis with the patient information unit limited in the same thread.</td>
</tr>
<tr>
<td>PA</td>
<td>The Patient-centered and Across-thread method: co-occurrence analysis with the patient information unit that considers across-thread association. It includes the co-occurrence of a drug and its potential adverse reaction that come from two different threads but associated with the same patient.</td>
</tr>
<tr>
<td>PIE</td>
<td>The PI method with Experience mining: the drug and ADR were experienced by the same patient in the same thread.</td>
</tr>
<tr>
<td>PAE</td>
<td>The PA method with Experience mining: the drug and ADR were experienced by the same patient but not limited in the same thread.</td>
</tr>
</tbody>
</table>
4.3.4 Evaluation with SIDER

We show the results for two most frequently mentioned drugs in the MedHelp heart disease forum: Aspirin and Atenolol, and one randomly selected drug: Diazepam, which was less frequently mentioned in this forum but also caused some adverse reactions for heart disease patients.

As we aim to discover potential ADRs for human experts to further verify, the task is essentially an ADR retrieval task and can be evaluated with the rank-based measures, such as the precision @$K$ and recall @$K$ used in [81]. We rank the ADRs based on the confidence defined in Eq. 4.2, and evaluate the precision and recall for the top-$K$ returned results. For a given drug, the precision @$K$ is the ratio of the number of correctly returned ADRs to the total number of returned ADRs given $K$. The recall @$K$ is the ratio of the number of correctly returned ADRs given $K$ to the total number of ADRs in the data set. For a given drug, we use the number of ADRs in SIDER to approximate the total number of ADRs in the data set. The reason is that it is too expensive for us to count the total number of ground-truth ADRs in the data set by manually checking if each pair of drug and adverse reaction is a ground-truth ADR. It is possible that the data set does not contain all the listed ADRs in SIDER. So the recall derived in this way is just proportional to the recall according to the above definition.

Fig. 4.1, Fig. 4.2, and Fig. 4.3 show the performance comparison of the variants of our proposed method for drug Aspirin, Atenolol, and Diazepam, respectively. According to Fig. 4.1(a), Fig. 4.2(a), and Fig. 4.3(a), the overall trend of the precision is decreasing with the increasing of $K$. When $K$ is sufficiently large to retrieve all the results, the precision becomes constant. In such case, we can clearly observe that $PI$ has a higher precision than $PA$, which indicates that the in-thread co-occurrence is
Figure 4.1: Top-K Evaluation with SIDER for Aspirin among the Variants of the Proposed Method

Figure 4.2: Top-K Evaluation with SIDER for Atenolol among Variants of the Proposed Method
more reliable than the across-thread co-occurrence in terms of ADR discovery. With a sufficiently large $K$, $PIE$ or $PAE$ has a much higher precision than $PI$ or $PA$, which confirms the positive effect of experience mining. Fig. 4.1(b), Fig. 4.2(b), and Fig. 4.3(b) show that by considering across-thread association, the recall can be significantly improved when $K$ becomes larger. A high recall is important if we want to identify some rare ADRs, which were mentioned less frequently and not necessarily mentioned by a patient in the same post or thread. Among all the proposed methods, $PAE$ achieves a good trade-off between the precision and recall, which will be recommended as our default method for further comparison.

Fig. 4.4, Fig. 4.5, and Fig. 4.6 show the performance comparison of our proposed method $PAE$ with the existing methods. Similarly, the overall trend of the precision is decreasing with the increasing of $K$. When $K$ is sufficiently large to retrieve all the results, the precision becomes constant. In such case, $Sent$ achieves the highest
Figure 4.4: Top-K Evaluation with SIDER for Aspirin by Comparing with the Existing Methods

(a) Precision@K

(b) Recall@K

Figure 4.5: Top-K Evaluation with SIDER for Atenolol by Comparing with the Existing Methods

(a) Precision@K

(b) Recall@K
Figure 4.6: Top-K Evaluation with SIDER for Diazepam by Comparing with the Existing Methods

precision, followed by Post, PAE, and then Thread. In Fig. 4.4(b), Fig. 4.5(b), and Fig. 4.6(b), when $K$ is relatively small, PAE achieves a similar recall to Thread and Post, which is better than Sent. With a sufficiently large $K$ to retrieve all the results, PAE achieves a similar or slightly lower recall than Thread, but better or much better than Post or Sent. Overall, PAE achieves a good trade-off between precision and recall. PAE also achieves a high precision when $K$ is relatively small. The reason is that PAE aggregates more information and retrieves candidate ADRs supported with higher co-occurrence frequency, which results in a more robust top-K performance. For all three drugs, the highest precision is achieved by PAE, which is 0.55, 0.55, and 0.28, given the top-20, top-20, and top-100 results, respectively.

We notice that the precision is not monotonically decreasing, especially when $K$ is relatively small. Such phenomena have also been observed in [81], in which the precision @10 is much higher than the precision @3. In our experiment, we observed
that the top ranked results do not guarantee that they are the most likely true ADRs. For example, the top-10 returned candidate ADRs often include some false positives like “stressed”, “worried”, “panic”, and “heart attack”, which can be an ADR of other drugs in SIDER. But in our cases they are mostly common feelings, symptoms, or diseases that the patient had. One possible solution to mitigate such problem will be discussed in the future work section. Some methods or scenarios, such as Sent or PIE for Diazepam, which retrieve candidate ADRs supported with lower co-occurrence frequency and get less reliable results, are more vulnerable to such noise. As we can see from Fig. 4.3(a), PIE achieves a low precision for Diazepam. For Diazepam, the returned candidate drug-ADR pairs were supported with low co-occurrence frequency, e.g., 18, 10, and 2 times for the 20th, 40th, 60th candidate ADR, respectively. They are less reliable compared with PIE for Aspirin, which are 37, 25, and 18 times, respectively, or PAE for Diazepam, which are 32, 22, 18, respectively. That also partially explained why PAE has a higher precision than PIE for Diazepam when K is small. More comparison between PIE and PAE will be discussed in the evaluation with human annotation.

4.3.5 Evaluation with Human Annotation

As observed in the existing work, there are some mismatches between the terms from the official knowledge base and those from the health forums, which justifies a need of evaluation with human annotation. In this work, we follow their procedure to design the human annotation experiments. Besides, we also want to investigate how the proposed system components, such as across-thread association and experience mining, directly affect the performance via the evaluation with human annotation. Two graduate students were hired to annotate the data. In the human annotation experiment, we focus on the ADRs for Aspirin, which is the most frequently mentioned
drug in the heart disease forum. We randomly select a subset of 130 threads that have mentioned Aspirin, which include 763 posts with 7859 sentences. Among them, we have identified 913 drug-ADR pairs in which the drug and ADR occur in the same thread. Among the 913 pairs, 221 pairs were labeled as ground truth, which means the same patient took the drug and experienced the adverse reaction, and we cannot obviously rule out the possible causal relationship between them.

We further checked the across-thread drug-ADR pairs. Here the goal is to identify the drug and adverse reaction that occur in two different threads but connected by the same patient. Specifically, we require that one thread mentions that the patient took the drug, and in another thread, the same patient experienced the adverse reaction. Unless we can easily rule out the causal relationship between the drug and the adverse reaction, we consider that the drug and the adverse reaction as a ground-truth pair. We identified 28 such pairs as the ground truth. Together, among the 130 selected threads, we have identified 249 drug-ADR pairs as ground truth.

Different from the large-scale evaluation with SIDER, now as we have identified all the ground-truth pairs in the sampled data set, we can use the set-based evaluation measures: precision ($P$), recall ($R$), and f-measure ($F_1$). The precision is the ratio of the number of correctly returned drug-ADR pairs to the total number of returned drug-ADR pairs. The recall is the ratio of the number of correctly returned drug-ADR pairs to the total number of drug-ADR pairs that should be returned according to the ground truth. The f-measure is defined as the harmonic mean of precision and recall: $F_1 = \frac{2PR}{P + R}$. 

71
Table 4.3: Evaluation with Human Annotation

<table>
<thead>
<tr>
<th>Method</th>
<th># of Correct Pairs Returned</th>
<th>Total # of Returned Pairs</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent</td>
<td>11</td>
<td>17</td>
<td>0.647</td>
<td>0.044</td>
<td>0.083</td>
</tr>
<tr>
<td>Post</td>
<td>120</td>
<td>211</td>
<td>0.569</td>
<td>0.482</td>
<td>0.522</td>
</tr>
<tr>
<td>Thread</td>
<td>221</td>
<td>941</td>
<td>0.235</td>
<td>0.888</td>
<td>0.371</td>
</tr>
<tr>
<td>PI</td>
<td>208</td>
<td>432</td>
<td>0.481</td>
<td>0.835</td>
<td>0.612</td>
</tr>
<tr>
<td>PA</td>
<td>236</td>
<td>1111</td>
<td>0.212</td>
<td>0.948</td>
<td>0.347</td>
</tr>
<tr>
<td>PIE</td>
<td>144</td>
<td>223</td>
<td>0.646</td>
<td>0.578</td>
<td>0.610</td>
</tr>
<tr>
<td>PAE</td>
<td>153</td>
<td>274</td>
<td>0.558</td>
<td>0.614</td>
<td>0.585</td>
</tr>
</tbody>
</table>

Table 4.3 shows the evaluation results with human annotation. Among the existing methods, *Sent* achieves the highest precision but with the lowest recall. *Thread* achieves the highest recall but with the lowest precision. Among our proposed methods, *PI*, *PIE*, and *PAE* outperform the existing methods in F1 measure. *PI* improves the precision over *Thread* with a relatively high recall, and thus achieves the best F1 performance, 0.612, followed by *PIE*, which has a very close F1 performance. *PA* achieves the highest recall, and it can also successfully retrieve all the 28 across-thread pairs. However, *PA* has the lowest precision. We investigated the data and found that many patients who participated in multiple discussion threads mentioned some adverse reactions not related to the drug taken by them. By including such drug-ADR pairs across threads, we introduced a large number of false positives, and decreased the precision of *PA*. Fortunately, with experience mining, *PAE* improves the precision and achieves a much better F1 performance.

We notice that some techniques that are integrated into our system are not perfect.
Table 4.4: False Positive Examples

<table>
<thead>
<tr>
<th></th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“I have had chest pains on and off for over four years....”</td>
</tr>
<tr>
<td></td>
<td>“About three weeks ago I got a really bad pain in my chest and went back to an emergency doctor.... The hospital watched me overnight, gave me aspirin over a few hours and told me to go home and....”</td>
</tr>
<tr>
<td>2</td>
<td>“Had Myocarditis from viral infection 2-years ago had some A-Fib 9-months later and great recovery from that after electrocardioversion....”</td>
</tr>
<tr>
<td></td>
<td>“Moved out of state since last year. New cardiologist in my new town placed me on Toprol 25 mg per day with 1- Aspirin and lisinopril 2.5mg per day....”</td>
</tr>
</tbody>
</table>

Table 4.5: False Negative Examples

<table>
<thead>
<tr>
<th></th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“i cannot tolerate one single aspirin, not even the coated ones, not even a baby aspirin, my stomach lining will be on fire.”</td>
</tr>
<tr>
<td>2</td>
<td>“The doctor didn’t seem to concerned, and just told me to go on one aspirin a day....the doctors seem to say that the palpitations aren’t life threatening, so to try to live with them.”</td>
</tr>
</tbody>
</table>

Consider the Stanford coreference resolution tool as an example, which achieves about 70% F1 performance in their report. It considers two person mentions, “Sharon” and “My daughter”, which occur in the same post and refer to the same person, as two different persons by mistake. The problem becomes even more serious due to the informal language used in a forum. For example, an acronym “my DD” can not be identified as the same person as mentioned in another post “my dear daughter”. Such problems can decrease the recall of our patient-centered method.

There are some limitations for our current system. We analyzed the false positives
and false negatives introduced by our system. First, as our methods are based on co-occurrence analysis, it is not necessary that the adverse reaction is caused by the drug. In Table 4.4, example 1 shows that the patient first had chest pain, and then took aspirin. Although aspirin can cause chest pain, here the patient’s chest pain is obviously not caused by aspirin. Similarly, in example 2, the patient’s viral infection cannot be caused by aspirin given their temporal order. In future, we plan to adopt some machine learning techniques, such as the probabilistic graphic models [81, 67], to identify the causal relationship to improve the ADR discovery performance. We also identified some other reasons for the false positives. For example, if we missed the identification of one patient, who experienced an ADR, the information of the unidentified patient may be merged with the information of another patient, who has taken the drug. In such case, the patient who took the drug is in fact not the same patient who experienced the ADR, which introduces a false positive.

For the false negatives, we first notice that there are some missed connections between two pieces of health information related to the same patient. For example, in the previous example, the Stanford coreference resolution tool cannot identify two person mentions “Sharon” and “My daughter” as the same person. In such case, if “Sharon” took the drug and “My daughter” experienced an adverse reaction, the drug and the adverse reaction cannot be considered as related, which introduces a false negative. We also notice that the recall of PAE is decreased due to the integration of experience mining. In other words, the experience mining introduced false negatives, which were true ADRs but missed by PAE. We investigated the reason and showed two false negative examples in Table 4.5. In Example 1 in Table 4.5, the sentence is labeled as not experience, which is very likely due to the use of modal verb or the verb in its future tense. However, example 1 in fact implies that the patient has taken the drug or experienced the adverse reaction. In example 2 in Table 4.5, the
same patient took Aspirin and experienced palpitations, which are a true ADR for Aspirin. However, since the major part of the sentence containing “palpitations” is about a doctor’s suggestion instead of a patient’s personal experience description, it is classified as not experience by our automatic experience mining component.

Overall, $PAE$ achieves a good balance between precision and recall for both evaluation with SIDER and human annotation. For evaluation with human annotation, $PAE$ outperforms all the three baseline methods in $F_1$ measure.

4.4 Summary

In this chapter, I presented the approaches and experiments about effectively discovering ADRs from health forums. The proposed patient-centered and experience-aware mining framework is used to build a patient health information database for effective ADR discovery. The experimental evaluation with both an official ADR knowledge base and a human-annotated ground truth verifies the effectiveness of the proposed approach for ADR discovery.
Chapter 5

RELATED WORK

This chapter briefly discusses the literature on mining, learning, and information discovery in health forums.

Mining health forums. There is a growing interest in mining online health forums. In [72], two popular machine learning methods: Support Vector Machines (SVM) and Conditional Random Fields (CRF), are used to classify the content of each sentence in forum posts into three types: symptom description, treatment description, or others. Such classification can be combined with our MetaMap or keyword list based methods to acquire more accurate health information labels. In [75], a probabilistic graphic model is used to establish the credibility of user statements in online health communities, which can be used to extract trustable drug side-effect statements. In [86], heterogeneous network mining techniques are used to detect drug-drug interactions between two given drugs. Opinion mining and sentiment analysis are also used for knowledge discovery in health forums [26, 14]. [26] discovers patient drug outcomes by clustering the forum topics and opinions. [14] compares the effectiveness of different treatments with sentiment analysis. The user’s demographic information is also used in [14] to differentiate the treatment effectiveness for different populations. Different from the existing work, we propose patient-centered information mining. It identifies the mentioned patients, aggregate, and associate health information in a forum post with the corresponding patients to form patient semantic information units for effective information discovery.
Adverse Drug Reaction Discovery. As shown in the latest survey papers [21, 34, 29], recently adverse drug reaction discovery in online health forums has received growing attention from different research communities. [11, 43, 44, 45, 85] study the problem of extracting or detecting ADR through association rule mining from health forums with different syntactic document units. It assumes that if a drug and an ADR term co-occur in the same syntactic unit (e.g., a sentence, a post, or a thread) frequently, then the drug is considered likely to cause that ADR. The most related work is from [43, 44, 45]. They developed a framework for pharmacovigilance in health social media. Their methods are focused on sentence-level analysis in which the mention of a drug and its potential adverse event co-occur in the same sentence. They also showed that manually distinguishing personal experience from hearsay can significantly improve the performance of identifying adverse drug events. Different from their work, we explored the ADR discovery based on patient semantic information units, and developed automatic patient experience mining into our system. In [11], the drug and the adverse reaction that co-occur within a window of 20 tokens in the same post are considered as potentially related. [87] is focused on ADR discovery in health forums based on thread-level analysis. They also use temporal analysis to identify ADR signals in their early stage. Some recent work uses probabilistic graphic models, such as topic models [81], hidden Markov models [67], and conditional random fields [88, 61], to extract ADRs from free text in health social media. They are either based on the analysis on a sentence, a post, or a thread information unit. They are orthogonal to our proposed methods, and can be integrated into our system in place of co-occurrence analysis in the future. [86] used the health forum data to mine the drug-drug interactions, which can be another potential application of our proposed methods.
**Person Resolution and Thread Structure Learning.** There are some research works related to person resolution on forums. First, some general coreference resolution or anaphora resolution systems can be used for person resolution [38, 37, 73]. However, although these general resolution systems are good at resolving mentions within a post, they are not suitable for coreference resolution across posts. The only coreference resolution system related to forums is introduced in [25], which focuses on the coreference resolution on blogs and commented news in Dutch. Blogs and commented news in Dutch are different from our health forums in English and their system is also not publicly available. Some coreference resolution or pronoun resolution systems for dialogues, such as those proposed in [55, 74], are also inspiring for our person resolution design on online forums, though they focus more on short-text dialogue in spoken language.

Recently, some research has been performed on learning or predicting thread structures for online forums, blogs, or news websites [40, 7, 79, 80, 77, 65]. In [40], topic modeling and temporal dependency between posts are incorporated in a sparse coding approach. In particular, one post is represented as a linear combination of all the preceding posts in the latent semantic space. The structure information is embedded by adding the constraints that the topics of each post can only be sampled from the topics of those preceding posts. The sparse coding approach can be used for reply relationship reconstruction, junk post detection, and expert finding. In [7], a classification approach is used to reconstruct the reply structure based on a set of simple features, such as time difference, content similarity, and quotation relationship. In [79], a joint classification approach with a linear-chain CRF or dependency parsing is used for predicting thread structures, which considers both the link relationships between posts and the dialogue acts assigned to each link. The dialogue acts are made from the five categories: question, answer, resolution, production, and other. In [80],
an unsupervised approach to predicting the thread reply structure is proposed, which utilizes the lexical chains between word tokens within a discourse to recover the inter-post links. Their work is orthogonal to ours and can be integrated into our learning framework. In [65], an extended block hidden Markov (HMM) model, which allows a state in HMM to correspond to a mixture of multiple classes, is used for unsupervised thread reply structure modeling. In [8], a supervised approach based on the ranking-SVM model is proposed to reconstruct the thread structures in blogs, online news agencies, and news websites, which are slightly different from online forums. In this work, we extend the threadCRF model [77] by considering the partially known structures and the abundant person reference information available in person-centric forums.

To address the partially labeled data problem, a semi-supervised training procedure for conditional random fields (CRFs) has been proposed in [27], which can be used with a combination of labeled and unlabeled training data. However, instead of having some instances fully labeled and some unlabeled, each training instance in our setting, which is a thread, is partially labeled. Therefore, the above semi-supervised training procedure is not suitable for our problem. In [76], a training procedure with incomplete annotated sentence instances is proposed for the Japanese word segmentation and part-of-speech tagging tasks. Instead of materializing sentence annotations in their application, we adopted a similar idea for training the threadCRF model with materialized thread structures.

**Experience Mining.** Some related work about mining experience from the Web includes [63, 28, 60, 43, 58]. [63] has been focused on domain-independent and objective experience mining from Weblogs. They proposed a set of sentence-level linguistic features to classify a sentence. [60] is focused on the detection of experience revealing sentences in product reviews. Similar to [63], each sentence is considered
independently. None of them use the context information to improve the sentence classification performance. In an explorative study, [28] investigates some features for automatic detection of reports of experiences with products from online forums. Although no further experiments are provided, these explored features can also be incorporated into our framework. [58] classifies a tweet instead of a sentence in a post in our work. The tweet context, such as the twitter’s profile, has been taken into consideration in their classification model, which is similar to our global context features. In [43], report source classification is proposed to differentiate patient experience from hearsay. In [44, 45], an evaluation based on 400 sentences manually labeled as patient experience or hearsay shows that differentiating them can significantly improve the adverse drug event extraction performance. In this work, we proposed a context-aware experience extraction framework and evaluated its effectiveness [50]. We also investigate the effect of automatic experience mining on ADR discovery.

Some work about experience knowledge mining is also related to our work. [4] is mainly focused on the factuality analysis of the event mentions described in a sentence, e.g., whether the event indeed took place. Although their proposed techniques are orthogonal to ours, they can be used to extract more sentence-level features to improve our patient experience extraction accuracy. [33, 59] focus on more fine-grained experience knowledge extraction, including the extraction of events, entities, and relationships.

**Patient-Centered Healthcare.** Recently, patient-centered healthcare or related research has attracted a lot of attention [70, 18, 1, 20, 9, 84, 68]. Patient-centered care has become a major aim for the national health system in the United States [70], which is supposed to replace the current physician centered system with one that revolves around the patient. In [70], six key components are proposed for patient-centered care: Education and shared knowledge; Involvement of family and friends;
Collaboration and team management; Sensitivity to nonmedical and spiritual dimensions of care; Respect for patient needs and preferences; Free flow and accessibility of information. The Patient-Centered Outcomes Research Institute (PCORI) is recently created by the Patient Protection and Affordable Care Act (ACA) to fund research that will provide patients, their caregivers, and clinicians with evidence-based information required to make better-informed health care decisions. Patient-Centered Outcomes Research (PCOR) is a relatively new research field that considers patients’ needs and preferences and focuses on outcomes most important to them. They try to answer patient-centered questions such as: “Given my personal characteristics, conditions, and preferences, what should I expect will happen to me?” “What are my options, and what are the potential benefits and harms of those options?” “What can I do to improve the outcomes that are most important to me?” “How can clinicians and the care delivery systems they work in help me make the best decisions about my health and health care?”

The proposed architectures or frameworks in [9, 84, 68] have confirmed the importance of patient-centered methods for information extraction, integration, retrieval, and knowledge discovery. In [9], information extraction from various knowledge bases, such as drug description repositories, medical encyclopedias, biomedical literature, and structured data from FAERS, has been integrated for patient-centered adverse drug event identification from patient ER records. In [84], a personalized health information retrieval architecture was presented, in which the profile data of a query user (a patient) is utilized to selectively retrieve the relevant medical information. In [68] an ontology-based visualization is presented, which can be used to analyze the patient’s profiles and the relationships between health data in the ontology knowledge base. Such ontology knowledge can also be integrated into our system for semantic information extraction and integration.
CONCLUSIONS AND FUTURE WORK

This chapter concludes the dissertation by summarizing the contributions of the work and discusses some potential future directions.

6.1 Conclusions

Online health forums provide a large repository for people to seek valuable information. A forum user can issue a keyword query to search health forums regarding to some specific questions, while a researcher can discover knowledge in a timely and large-scale fashion by automatically aggregating the latest evidences emerging in health forums. This dissertation studies how to effectively discover information in health forums. Several challenges have been identified and addressed.

First, the existing work relies on the syntactic information units, such as a sentence, a post, or a thread, to bind different pieces of information in a forum. However, most of information discovery tasks should be based on the semantic information unit, a patient. In this work, patient-centered mining is proposed to mine patient semantic information units for effective information discovery. In a patient information unit, the health information, such as diseases, symptoms, treatments, effects, and etc., is connected by the corresponding patient. As the thread structures, the reply relationships between posts, are very important for patient-centered mining, but most online forums only have partially labeled structures. A unified framework is proposed to learn the complete thread structures based on a statistical machine learning model, which leverage the existing partially known structures and the abundant person reference relationships in health forums. The effectiveness of the proposed approaches
for thread structure learning has been verified on two health forum data sets. The effectiveness of the patient-centered mining for effective information discovery has been verified with the proposed keyword query experiments.

Second, the information published in health forums has varying degree of quality. Some information includes patient-reported personal health experience, while others can be hearsay. In this work, a context-aware experience extraction framework is proposed to mine patient-reported personal health experience, which can be used for evidence-based knowledge discovery or finding patients with similar experience. The extraction of patient experience is modeled as a classification problem: classifying each sentence in a forum post as containing patient experience or not containing patient experience. The sentence context information is exploited for such experience extraction task. The context information is first classified into two types: global context and local context, and then incorporated into a unified context-aware experience extraction framework CARE. The experimental results show that the global context can significantly improve the experience extraction accuracy, while the local context can also improve the performance when less labeled data is available.

At last, the proposed patient-centered and experience-aware mining framework is used to build a patient health information database for effectively discovering ADRs from health forums. The experimental evaluation with both an official ADR knowledge base and a human-annotated ground truth verifies the effectiveness of the proposed approach for ADR discovery.

6.2 Future Work

In this section we discuss possible future works in several categories.

**Deep Learning for Text Mining.** Deep learning has recently shown much promise for NLP and text mining applications [71]. Traditionally, documents or sentences are
represented by a sparse bag-of-words representation. Recently, a distributed representation of words, such as “neural embedding” or vector space representation of each word or document, has been used for various NLP and text mining applications, such as sentiment analysis, machine translation, natural language inference, and etc. With deep learning, instead of manually crafting features, we can leverage unsupervised feature learning techniques [36, 5] to automatically learn good feature representations from big data for patient experience mining and ADR discovery. Such techniques are expected to be more robust and achieve better performance.

**Relation Classification.** One potential improvement of the current system is to further classify the relationship between the connected information within the same patient information unit. Consider the ADR discovery as an example. As our proposed methods mainly use co-occurrence analysis, some drug indications or other drugs’ adverse reactions can be identified as the ADRs of the target drug by mistake. For example, a patient may take a drug for an unapproved indication, i.e., a disorder treated by a drug that is not approved by FDA. A patient can also take multiple drugs and the ADR is probably caused by one of them or an interaction among two or more drugs. To identify which drug has caused the adverse reaction, we can adopt some machine learning techniques, such as the probabilistic graphic models, to identify the causal relationship, and thus improve the ADR discovery performance.

**Learning to Rank.** One possible future research direction is learning to rank the discovered information based on relevance and importance. Consider the ADR discovery as an example. We observed that the in-thread or experienced co-occurrences are more important for ADR discovery. We can thus assign them with a higher weight or rank them higher during the top-K ADR discovery process. Here the challenge is how to set these weights or rank. We can learn an effective ranking function from user studies [42].
**Integrating Multiple Sources.** In addition to online health forums, we would like to extend the proposed work to other social media that also contains valuable health information, such as Twitter, microblogs, and etc. Besides mining information from unstructured or semi-structured data, we can also integrate structured databases, such as the structured data from FEARS [2], into our system. Besides the user generated data, we can also integrate information from various knowledge repositories, such as drug description repositories, medical encyclopedias, biomedical literature, and etc.

**Crowdsourcing Evaluation.** As obtaining an accurate ground truth for evaluating our methods is critical but extremely labor-intensive, we will investigate obtaining ground truth through crowdsourcing [22], where the challenge is how to design tasks for the crowd and how to consolidate their opinions to obtain ground truth.
REFERENCES


