

Title	ラベルなしデータからの医学テキストマイニングのための遠距離教師あり学習とトランスダクティブ推定
Author(s)	Taewijit, Siriwon
Citation	
Issue Date	2017-12
Type	Thesis or Dissertation
Text version	ETD
URL	<a href="http://hdl.handle.net/10119/15072">http://hdl.handle.net/10119/15072</a>
Rights	
Description	Supervisor:池田 満, 知識科学研究科, 博士

# **Distant Supervision and Transductive Inference from Unlabeled Data for Medical Text Mining**

Siriwon TAEWIJIT

Japan Advanced Institute of Science and Technology



# Doctoral Dissertation

## Distant Supervision and Transductive Inference from Unlabeled Data for Medical Text Mining

by

Siriwon TAEWIJIT

*Supervisor:* Professor Mitsuru IKEDA

*School of Knowledge Science  
Japan Advanced Institute of Science and Technology*

December, 2017





# Abstract

Text mining has been increasingly significant due to the exponential growth of data. Dealing with text mining, the text preprocessing is the additional task to transform unstructured textual data into structured one. This machine-readable format benefits not only for data comprehension, interpretation, visualization but also further utilization by traditional data mining process. Notably, relation extraction is one of the keys for discovering hidden knowledge underneath the large-scale text. The relation extraction can be thought as the classification problem of a predefined relation from a given associative couple entities, for instance, the entity pair *Amoxicillin-diarrhea* and its relation *adverse reaction*. There are remaining key challenges and require manipulation by an efficient method. The supervised learning model is a well-known solution to learn attributes of a pair of entities and assigns a relation. However, the model accuracy is limited by the number of training examples, and the hand-labeled data acquisition from a large volume of text is also impracticable. While the extracting relation by pattern-based method is efficient, the manual processes for pattern generation and pattern selection are the restrictions. Moreover, the intractable processing of noisy, ill-form, domain-specific textual data and uncontrollable of unlabeled one are very challenging as well.

This dissertation mainly aims to cope with incomplete textual data (missing label) and improve the performance of relation extraction method. Regarding big data era, knowledge bases are reliable, freely available, inexpensive and maintained in multiple domains, e.g., Wikipedia for *person-organization* relation, SIDER for *drug-event* relation, IntAct for *protein-protein interaction* relation. The leveraging an existing knowledge base instead of manual label tagging can be seen as the promise solution for training data preparation or pattern generation, e.g., distantly supervised relation extraction by Freebase and Wikipedia. Additionally, a key phrasal pattern is a simplified version of a given sentence but retains a semantic, e.g.,  $\langle drug \rangle$ , *was-held-due-to*,

$\langle event \rangle$  is the phrasal pattern of the sentence “*On arrival here, propofol<sub>drug</sub> was held due to hypotension<sub>event</sub>*”. The word independence assumption is widely used in Naïve Bayes for text classification due to simple but effective, although, the current word is conditionally dependent on the previous word as shown in natural language. The key phrasal pattern can benefit to reduce model complexity in the dependency representation with three elements (a drug, a pattern, an event) for all sentences instead of length  $l$  of a given sentence. Using the appropriate assumption with such data representation can yield improvement in the classification model.

To this end, the dissertation presents a framework for relation extraction from unstructured text, and the medical text will be used as a case study to extract drug-event relation. Furthermore, the dissertation introduces parameters estimation in a generative model that argues word independence assumption. This contribution can dramatically improve a model performance. Lastly, the dissertation contributes the examination on multiple approaches of incomplete data incorporation for handling unlabeled data with the efficient way.

**Keywords:** adverse drug reaction (ADR), medical text mining, distant supervision, multiple-instance learning (MIL), relation extraction, transductive inference

# Acknowledgments

This dissertation would not have been completed without the help, support, suggestions, guidance, and effort of a lot of people. It gives me great pleasure to express my sincere thanks to whom I am greatly indebted.

Firstly, I owe my deepest gratitude to my supervisors, Professor Thanaruk Theeramunkong (Sirindhorn International Institute of Technology, Thammasat University), as well as, Professor Mitsuru Ikeda (JAIST) for their kindly guidance and support a lot of things since the first day when I am a fresh Ph.D. student to the present day when I am going to graduate.

I would like to express the appreciation to Dr.Sewan Theeramunkong (Pharmacy, Thammasat University), Dr.Ithiphan Methaseth (The National Electronics and Computer Technology Center - NECTEC), Professor Kenji Araki (MD, University of Miyazaki Hospital) and Mr.Peravas Pattanaprayoonwong (Pharmacy, Chulalongkorn University) who commented and suggested on the general questions relevant to pharmaceutical and medical domains and have provided evaluations on experimental results.

I would like to show my gratitude to my committee chairs, Professor Michitaka Kosaka, Professor Riichiro Mizoguchi, Associate Professor Van-Nam HUYNH and Associate Professor Dam Hieu Chi for their valuable comments and suggestions on my dissertation.

I would like to express my sincere thanks to JAIST-SIIT-NECTEC Dual Degree Program for providing me an opportunity to join the excellent research environment at JAIST and SIIT.

A very special thank to Dr.Ryosuke Matsuo (Knowledge Science, JAIST), Mr.Nuttapong Sanglerdsinlapachai, Mr.Saranyoo Sorkamnerd, members of Ikeda laboratory, Thais students in JAIST and members of Ho laboratory who shared their research ideas, life experiences and useful discussions including helped me to overcome tough times in my study at JAIST.

Finally, I dedicate this dissertation to my parents and family who always believed in me, encourage and support throughout my life.

# Table of Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>Table of Contents</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Relation Extraction from Clinical Text . . . . .	2
1.2 Research Problems . . . . .	4
1.3 Overview of Contributions . . . . .	6
1.4 Dissertation Outline . . . . .	7
<b>2 Background</b>	<b>8</b>
2.1 Text Mining . . . . .	8
2.2 Data Sources for Medical Relation Extraction . . . . .	11
2.3 Named Entity Recognition . . . . .	26
2.4 Open Information Extraction . . . . .	29
2.5 Distant Supervision . . . . .	30
<b>3 Relation Extraction in Clinical Text</b>	<b>33</b>
3.1 Drug-Event Relation Extraction . . . . .	33
3.2 Data Preparation . . . . .	41
3.2.1 Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III)	41
3.2.2 Sentence Boundary Detection . . . . .	46
3.2.3 Medical Named Entity Recognition and Normalization . . . . .	47

3.3	Evaluation Metrics . . . . .	52
3.3.1	Evaluation Metric for Drug-Event Association Analysis . . . . .	52
3.3.2	Evaluation Metric for Drug-Event Identification . . . . .	52
<b>4</b>	<b>Distant Supervision-Based Pattern Bootstrapping for Relation Ex- traction</b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	Idea and Contributions . . . . .	58
4.3	Proposed Method . . . . .	61
4.3.1	Distantly Supervised Initial Seed . . . . .	62
4.3.2	Automatic Phrasal Pattern Generation . . . . .	64
4.3.3	Phrasal Pattern Scoring . . . . .	66
4.3.4	Iterative Seed Generation . . . . .	67
4.3.5	Semantic Relation Inference . . . . .	67
4.4	Evaluation . . . . .	69
4.4.1	Data . . . . .	69
4.4.2	Analysis of Extracted Key Phrasal Patterns . . . . .	70
4.4.3	Evaluation on the Key Phrasal Pattern VS. Semantic Relation Specificity . . . . .	71
4.4.4	Evaluation on the Discovered Drug-Event Pair by Domain Experts	74
4.5	Summary . . . . .	79
<b>5</b>	<b>Distant Supervision-Based Transductive Inference for Relation Ex- traction</b>	<b>81</b>
5.1	Introduction . . . . .	81
5.2	Problem Formulation . . . . .	86
5.2.1	Distant Supervision . . . . .	86
5.2.2	Multiple Instance Learning . . . . .	87
5.3	Proposed Method . . . . .	88
5.3.1	Distantly Supervised Ground Truth . . . . .	91
5.3.2	Document Representation . . . . .	93
5.3.3	Transductive Learning for Relation Extraction . . . . .	96

5.3.4	The Incorporation of Unlabeled Data . . . . .	101
5.4	Evaluation . . . . .	104
5.4.1	Data . . . . .	104
5.4.2	Key phrasal patterns analysis . . . . .	106
5.4.3	Evaluation on the Effectiveness of the Key Phrasal Pattern-Based Feature . . . . .	107
5.4.4	Evaluation on the Effectiveness of MIL-dEM-SL and MIL-dEM-T	115
5.4.5	Evaluation on Overall Performance with Advanced Machine Learn- ing Methods . . . . .	116
5.4.6	Evaluation on Effect of Unlabeled Data Incorporation . . . . .	118
5.5	Summary . . . . .	120
<b>6</b>	<b>Conclusion</b>	<b>123</b>
	<b>Bibliography</b>	<b>125</b>
	<b>Publications</b>	<b>142</b>

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Sirindhorn International Institute of Science and Technology, Thammasat University.

# List of Figures

2.1	A taxonomy of text preprocessing tasks . . . . .	9
2.2	Text mining processes: from unstructured data to structured content. .	10
2.3	Example of heterogeneous data in EMR. (a) Textual data in a discharge summary. (b) Magnetic Resonance Angiography (MRA) images. (c) ECG signal. . . . .	14
2.4	An example form of spontaneous report in Med Watch. . . . .	20
2.5	An example data from DrugBank. . . . .	22
2.6	An example of MEDLINE data in XML format. . . . .	24
2.7	An example of IEEE Xplore API data in XML format. . . . .	25
2.8	An example of Tweets data. . . . .	27
3.1	MIMIC-III: a portion of event notes per category . . . . .	42
3.2	MIMIC-III: Total number of sentences over sections in discharge summary	47
3.3	The phenotype of atrophoderma vermiculatum disorder . . . . .	49
4.1	The bootstrapping method for drug-event relation extraction. . . . .	61
4.2	The proposed method for phrasal pattern bootstrapping for semantic relations identification. . . . .	62
4.3	Open Information Extraction for given two medical sentences from EMR. . . . .	66
4.4	Plot of discovered key phrasal patterns across frequency and pattern strength. . . . .	69
5.1	Overview of the proposed adverse drug reaction identification framework	90
5.2	Medical named entity recognition and relation candidate generation. . .	92

5.3	Block-1 distant supervise for automatic data labeling. Block-2 depicts the proposed MIL-dEM method. . . . .	93
5.4	The key phrasal pattern plot by adjusted entropy score vs. frequency.	106
5.5	The number of features for each type of pattern weighting method across F1-score. . . . .	110
5.6	F1-score vs. numbers of unlabeled data . . . . .	120

# List of Tables

2.1	The characteristic of data sources in biomedical domain. . . . .	12
2.2	The electronic medical record data sources for medical text mining. . .	15
3.1	A list of previous studies on drug-event relation extraction . . . . .	35
3.2	Statistical number of note event categories from EMR and example of sentences. . . . .	43
3.3	(a) An example of a narrative note from a discharge summary in EMR system. (b) The noise-prone from a given text. . . . .	48
3.4	An example of a partial narrative notes from MIMIC-III. The drug and event entities are expressed in bold. . . . .	49
3.5	An example of partial narrative notes from MIMIC-III. The medical terms (in bold) present the same disorder. . . . .	50
3.6	The statistical number of narrative notes from MIMIC-III after data preparation. . . . .	53
4.1	The statistical number of extracted relations derived by OpenIE from narrative notes in MIMIC-III. . . . .	65
4.2	Top 5 of the extracted key phrasal pattern and the example sentences from MIMIC-III. . . . .	68
4.3	The summary table of the key phrasal pattern comparison. . . . .	72
4.4	The comparison of the key phrasal patterns between the proposed method and the two studies of Xu et al. . . . .	73
4.5	The evaluation on key phrasal patterns: discovered drug-event pairs vs. knowledge base. . . . .	75
4.6	The evaluation by domain experts on the randomly selected sentences.	77

5.1	A list of previous studies on ADR identification from unstructured text	84
5.2	Types of feature extraction for a given sentence. . . . .	95
5.3	The list of parameters for assessment . . . . .	105
5.4	An example of relevant sentences of drug-event (d, p, e) pairs. . . . .	108
5.5	The effectiveness comparison on 5-fold cross validation of text representation across three types of document weighting using MIL-iEM with <i>soft decision making</i> (MIL-iEM-S). . . . .	111
5.6	The effectiveness comparison on 5-fold cross validation of text representation across three types of document weighting using MIL-iEM with <i>hard decision making</i> (MIL-iEM-H). . . . .	113
5.7	The effectiveness of MIL-dEM-S-SL and MIL-dEM-S-T comparison across three types of initial weight on 5-fold cross validation with <i>soft decision making</i> . . . . .	117
5.8	The comparison of overall performance among MIL-dEM-SL, MIL-dEM-T, advanced machine learning methods, and MIL-iEM-T using 5-fold cross validation. . . . .	119

# List of Abbreviations

ADR	Adverse Drug Reaction
B	Binomial Distribution
BOW	Bag-of-Words
CUI	UMLS Concept Unique Identifier
dEM	Expectation-Maximization with dependency representation
EM	Expectation-Maximization
EMR	Electronic Medical Record
iEM	Expectation-Maximization with independent assumption
IND	Indication
MIL	Multiple-Instance Learning
MILR	Multiple-Instance Logistic Regression
MINB	Multiple-Instance Naïve Bayes
MISVM	Multiple-Instance Support Vector Machines
NB	Naïve Bayes
OpenIE	Open Information Extraction
SRS	Spontaneous Reporting System
TF	Term Frequency
TFIDF	Term Frequency-Inverse Document Frequency
TSVM	Transductive Support Vector Machines
UMLS	Unified Medical Language System

# Chapter 1

## Introduction

Like as image, body language or sound, a text is written in symbolic language for communication purpose, and moreover, a text uses the specifically written symbols to convey a message with more explicit meaning. This written communication is well-known as the most common and very effective method for information transmission and storing. Text for communication might be expressed as a conceptual or a particular meaning range from a one word such as “headache” to multiple words such as a phrase “abdominal pain” or a sentence “propofol causes mild hypotension.” and so on.

Given a text, human has expertise in understanding, reasoning, summarizing, interpreting, even inferring whether the content is short or long text, complete or incomplete sentence, with or without noise. Such cognitive process is computed through the complexity of a human nervous system. However, human intelligence has the capacity limitation in information processing on an amount of data at a time. It motivates artificial intelligence and machine learning researchers to imitate human abilities as an automated computational model to deal with massive amount of textual data, especially in the era of “Big Data”.

While textual data is ubiquitous, textual information is rare and not readily available. Text mining is a key to successful knowledge extraction from a document collection over time. The distinctive characteristics of text that brings challenging to text mining are unstructured (or semi-structured) data, amorphous, and contains information at many different levels [1]; further, the difficulty is increased not only rely on

a domain of interest but also language dependency. Discovering information in free text, so-called *Information Extraction* is expected to capture a real-world object, i.e., people, place, organization, location, time with its attributes mentioned in a text [2] or with a relation between a pair of entities.

Relation extraction is well-known as a subtask under information extraction. At least two main tasks are involved in relation extraction; relation candidate generation and relation classification. The former includes the identification of a link (connection) between entity pairs. However, such link without relation label seems to lose essential information regarding semantic comprehension, e.g., *Steve Jobs-Steve Wozniak* relation has appeared as a couple of nodes with a link in a people-people network, but what is kind of relationship between such a couple successful guys. The above-mentioned question can be obtained by the latter task. The relation classification aims to assign relation label to a link of a couple entities, e.g., the relation label of *Steve Jobs-Steve Wozniak* relation can be *co-founder* of Apple company.

## 1.1 Relation Extraction from Clinical Text

Recently, due to the massive growth of electronic medical record (EMR), EMR repository is recognized as a promising source of data to discovering hidden patterns or association. EMR contains a collection of tacit knowledge [3]—professionals’ experiences, know-how, and intuitions; and explicit knowledge—scientific research literature, disease diagnosis procedure, patient information; in the form of a digital version of structured and unstructured texts. This repository offers insight into significant towards healthcare problems, e.g., patient mortality prediction [4], patient risk identification [5, 6], drug-disease relation extraction [7], drug-drug interaction prediction [8, 9]. Accordingly, the changing from paper-based to electronic records has brought to challenges and opportunities toward clinical text mining, e.g., text classification, text summarization, and information extraction.

Particularly, one of the potential medical applications is Adverse Drug Reaction (ADR) surveillance. The ADR terminology is defined as “*response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis,*

*diagnosis or therapy of disease, or for modification of physiological function*” [10]. Basically, computational approaches for drug safety surveillance can be classified into two categories regarding the difference of problems; ADR prediction and ADR identification [11]. The former approach is mostly involved in *pre-market surveillance*, and the latter one can be found in *post-market surveillance*.

- **ADR prediction** aims to construct a model for predicting unknown ADRs that have never been reported in anywhere.
- **ADR identification** targets on the retrieval process of existing ADRs in a given set of historical data, but they are not explicitly described as knowledge. The identifying process can be achieved using the data-driven with collected data from an experience of patient drug usage.

Regarding text mining approach, unstructured text from EMR is the promising source for ADR identification. In an earlier study on ADR identification, the statistical analysis of words co-occurring within a give clinical sentence is broadly employed to quantify the relationship strength between a couple of drug and clinical event, so-called *drug-event* pair. Unfortunately, the statistical association method exhibits the major pitfall. A discovered drugevent pair might not express clinical relevance [12] due to ignorance of relational context analysis. The surrounding context around mentioned drug-event pair within a given sentence can reveal an exact impressive in a clinical event, e.g., adverse drug reaction (ADR) or therapeutic indication (IND). The widely used supervised learning method encounters with a rare availability of training data (labeled examples) even though many unlabeled instances may exist. Toward this insufficiency of labeled examples but plenty of unlabeled ones, the previous studies of many researchers exhibit to assign labels of unlabeled examples either using a model learned from labeled examples (semi-supervised learning [13]) or using a sort of heuristics or rules (distant supervision [14, 15]). Then later such instances the labels of which are assigned are included in the training dataset for further model revision.

According to model categories in semi-supervised learning, three parameters are (i) predictive model, (ii) single model or collaborative model and (iii) test instances

handling model. As the first parameter, recent works [16, 17, 18] have proposed various models, such as generative models [19, 20], low-density separation models [21], and graph-based models [22]. For the second parameter, at least two alternatives, namely self-training [23, 24] and co-training [25], can be applied to assign a label to an unlabeled example by either one single predictive model or multiple ensemble classification models. The last parameter concerns with how to handle examples in validation dataset, where two choices are (i) to manipulate validation examples separately from unlabeled examples (inductive learning) or (ii) to treat such validation examples as unlabeled examples in the training step (transductive learning). Regardless of any choice in those as mentioned above, semi-supervised learning requires some labeled examples for an initial model construction, bringing the complexity in the acquisition of such initial labeled data same as shown in supervised learning method. As a recent solution, distant supervision has been proposed as a method to obtain a labeled training dataset that each unlabeled example is automatically assigned the most suitable label by means of heuristics or rules.

## 1.2 Research Problems

*“—Bringing together knowledge science  
and computational method to innovate novel solution —”*

Knowledge science provides metacognition, which is the important beginning process to solve a particular problem. On the one hand, a computational machine learning method is an intelligence tool that can aid knowledge discovery task effectively. The key question is how to convey the metacognitive skill from human to machine?

In this dissertation, I study relation extraction to discover hidden knowledge of associative drug and event pair that expresses relevance clinically from unstructured clinical textual data. Accordingly, I address three fundamental problems

- (i) The impracticable for hand-labeled examples.
- (ii) The intractable processing of unstructured text.

(iii) The effect from uncontrollable of unlabeled data to model robustness.

For the first problem, the lack of domain experts for examples labeling task is a fundamental issue when deals with either supervised learning or semi-supervised learning method. The model accuracy is limited by a small number of training examples as well. Even though the extracting relation by a pattern-based method is efficient, the manual processes for pattern generation and pattern selection are the restrictions. In the large-scale textual data, the process of hand-labeled examples is well-known as tedious and laboring tasks. The cost and time-consuming of such manual label tagging are linearly expensive along with number of class labels. It is infeasible for fewer efforts when shifting to a new class label, and such tagging labels process is repeatedly required.

The second problem is introduced by noisy, ill-form and domain-specific textual data that bring challenging to data preprocessing and feature engineering processes. The bag-of-words model corresponding word independence assumption is widely used in Naïve Bayes for text classification including relation extraction due to simple but effective. While a current word is conditionally dependent on the previous word as found in the typical natural language, the dependency representation model is oblivious due to needed many numbers of parameters estimation. The appropriate sentence representation combining with the dependency assumption among consecutive words can yield improvement in the classification model.

Lastly, due to the availability of vast amounts of textual data, it is well-known that labeled data is limited, rare, or expensive, while unlabeled data is much cheaper and freely accessible. Even though the supervised learning model is efficient, the model has some draw bank regarding a size of labeled data for learning model. A small number of labeled data but plenty of unlabeled ones can hurt the model performance. Therefore a robust classification model is required for uncontrollable of such unlabeled data.

In summary, to deal with the dissertation topic, firstly the process focuses on how to create new knowledge from an existing knowledge; the metacognitive skill is conveyed from human to machine. Then machine learning imitates the human process. The novel solution is served to tackle a technical problem. Finally, the newly discovered

knowledge is feedback into the existing knowledge and used to gradually discover new knowledge further.

## 1.3 Overview of Contributions

This dissertation contributes into two different viewpoints.

(i) Academic viewpoints:

- The dissertation presents a framework as a solution for semantic relation extraction from textual data. Moreover, a framework can also be applied in any domain with certain assumption.
- The dissertation introduces alternative parameters estimation in a generative model that provides improvement of model performance than the traditional word independence assumption. This contribution can help to improve the performance of a model dramatically.
- The dissertation contributes to the examination of the multiple approaches of unlabeled data augmentation to deal with uncontrollable of large-scale unlabeled textual data with the effectiveness.

(ii) Social viewpoints:

- The application of this research can support the health and safety surveillance from a drug usage
- The framework can help to suggest the possible harmful and beneficial relations to a professional healthcare and might bring discovery into the medical domain

In order to claim the above evidence, the dissertation exhibits throughout the multiple experiments.

- The experiment of the effectiveness of feature extraction: The dissertation expresses the examination of proposed feature in two viewpoints: (i) the ability

of the feature to identify ADR; (ii) the performance of the feature compares to baseline feature such as bag of word

- The experiment of the effectiveness of alternative parameters estimation method in a generative model: The experiment is conducted by the assessment of a proposed method and various advanced methods.
- The experiment of the effectiveness of unlabeled data incorporation to deal with large-scale textual data: The experiment investigates the robustness of the proposed model by varying the size of unlabeled data augmentation.
- The experiment of the effectiveness of the proposed framework in practical application: The domain expert is invited to be involved in data validation including commentary on experimental results.

## 1.4 Dissertation Outline

The dissertation is organized into five chapters, as follows:

Chapter 1 introduces the research problem and its formulation. This chapter also states the main contributions in term of biomedical findings and computational methods.

Chapter 2 presents the background of the dissertation. The studies of text mining, particularly, the medical domain is discussed.

Chapter 3 exhibits the common pre-processing method, data and evaluation that used in the dissertation.

Chapter 4 describes the proposed computational method for identifying adverse drug reaction by distant supervision and bootstrapping approach.

Chapter 5 presents the proposed method to enlarge a small number of training data for ADR identification by distant supervision and a generative model.

Chapter 6 concludes the dissertation by summarizing the major contributions, achievements, and limitations of the work. The future research on this topic also discussed in this chapter.

# Chapter 2

## Background

### 2.1 Text Mining

Given unstructured or semi-structured textual data in large collections of a document, the definition of knowledge discovery from text [26] or text mining is the process for the sake of extracting useful information and non-trivial patterns or knowledge from such textual data [27] through the identification and exploration of interesting patterns [28]. Text mining can be accomplished by either handcrafted method [29, 30, 31], automatic method [32] or hybrid method [33, 34]. Certainly, text mining is inspired by data mining in which known as knowledge discovery in databases. Therefore, one dominant difference between data mining and text mining is relevant to preprocessing process. While text mining requires more processing for the identification and extraction of representative features for text written in a natural language to transform unstructured textual data into a more explicitly structured intermediate format, data mining is not relevant to such process.

#### **From Text to Knowledge**

The preprocessing task in text mining aims to simplify textual data in a new form of machine-readable to enable by traditional data mining or machine learning methods. According to [28], a taxonomy of text preprocessing tasks is depicted in Figure 2.1.

Typically, information extraction aims to extract hidden information from large-

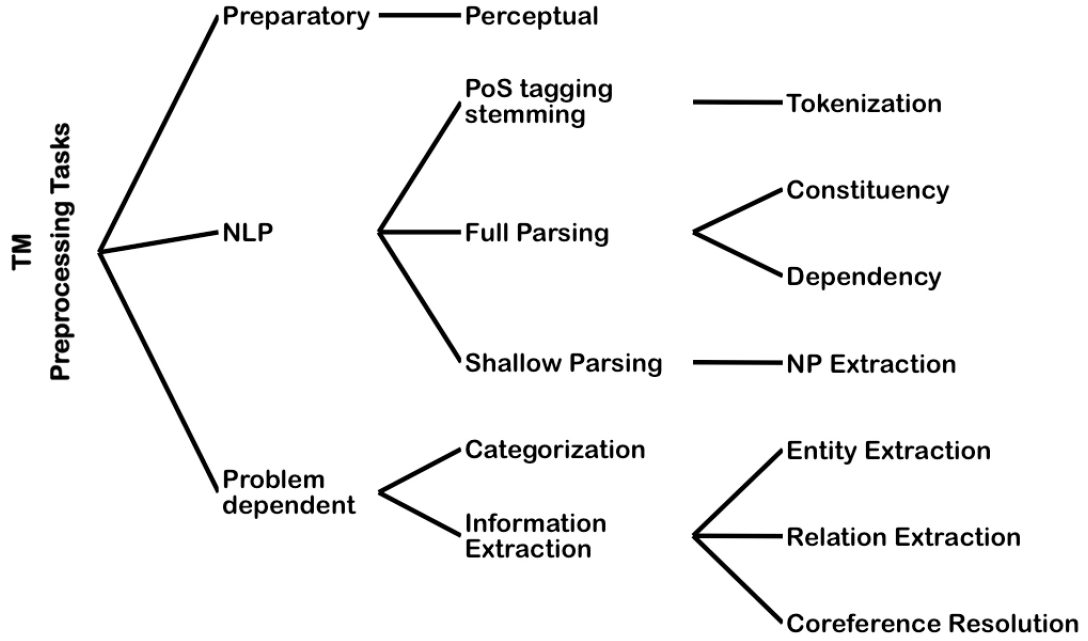


Fig. 2.1: A taxonomy of text preprocessing tasks

scale of textual data. As a formal definition, information extraction can be thought as “a process of getting structured information from unstructured data in text” [35]. For more specific, Grishman [36] gives a definition of information extraction as a process for “the identification of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship”.

Figure 2.2 exhibits the related components in order to identify, extract, and transfer implicit knowledge from unstructured textual data to explicit knowledge that can convenience for further utilization through structured data of text in the form of graph network or table in a database.

- **Information Retrieval:** the finding of relevant documents containing answers to a specific question, but not the finding of answer itself to such question [37]. The process is normally accomplished by means of statistical measures for information representation and document comparison [38]. The output of this process is a set of documents.
- **Information Extraction:** the extracting of specific information such as name,

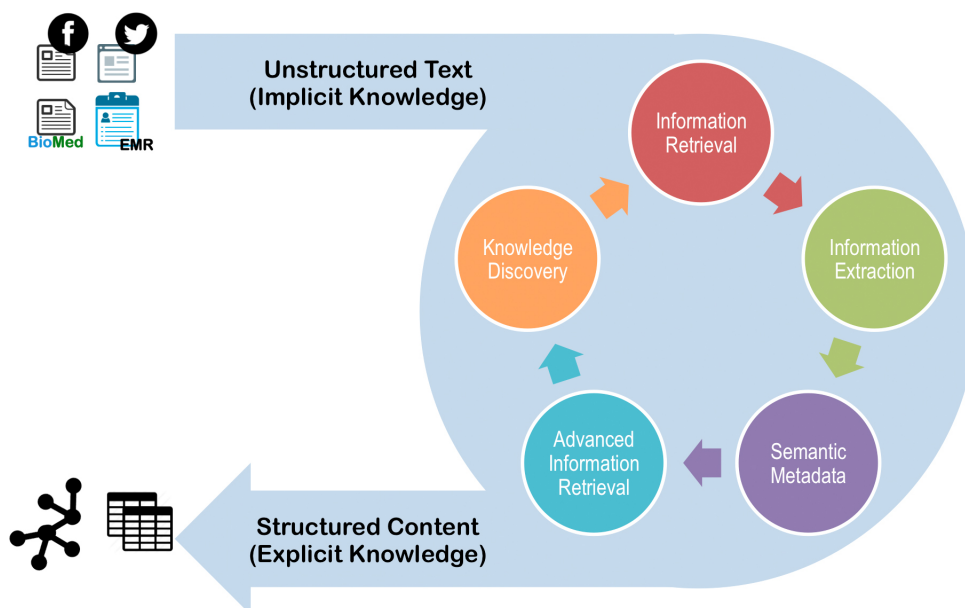


Fig. 2.2: Text mining processes: from unstructured data to structured content.

pattern or relation from given texts. The output of this process can be seen as structured data or network graph.

- **Semantic Metadata:** the quality improvement of the extracted information by considering a meaning of an individual text or a set of text at the conceptual level. The output of this process can be thought as words synonym or a group of similar words or phrases.
- **Knowledge Discovery:** the identifying of valid, non-trivial, novel, potentially useful, and ultimately understandable patterns in data. The output of this process can be new knowledge, new pattern or new relation.

As the case study on medical relation extraction, this dissertation aims to extract the relationship between prescribed drugs and observed clinical events unstructured text in EMR. The relevant computational methods to text mining technique are developed such as sentence boundary detection, named entity recognition, relation detection, and relation classification.

## 2.2 Data Sources for Medical Relation Extraction

Data collection is recognized as a key element in text mining. The different type of data sources could provide either many challenges or dissimilar problem-solving method. The source of data can be ranged from a reliability and well-written form (e.g. clinical narrative text, literature, encyclopedia) to noise-prone (or ill-form) and less trustworthiness source of data (e.g. blog, web forum, twitter). For medical relation extraction, textual data can be obtained by EMR system, spontaneous reporting system (SRS), biomedical literature, omics database or social media (see Table 2.1).

### Electronic Medical Record (EMR)

Earlier, EMR system is designed for the sake of warehouse system to record all patient dimensions into electronic format [40, 41]. This repository contains a collection of tacit knowledge [3] (e.g., professionals' experiences, know-how) and explicit knowledge (e.g., diagnosis procedure, patient information) in a digital form of structured and unstructured data. Moreover, EMR repository also offers insight into significant healthcare problems: patient mortality prediction [4], patient risk identification [5, 6], drug-disease relation extraction [7], drug-drug interaction prediction [8, 9] adverse drug reaction detection [12, 42]. There are numerous advantages of making use of EMR; (i) the high-reliability repository regarding terminology, controlled vocabulary and nomenclature code; (ii) to facilitate relational structure for effortless data acquisition (iii) longitudinal patient care and outcome; (iv) flourishing positive and negative patients risks observations to assess the safety and efficacy of a drug. The favorable merit of EMR differs from other data sources mentioned in Table 2.1 in the sense that those data sources lack in clinical sensibility and some of them are less trustworthiness. In addition, medical literature, spontaneous report ordinary fall into biased data corresponding to only passive outcome monitoring [28]. These overwhelmed drawbacks make extremely intriguing to recent researches.

Among massive of EMR data, doctor daily notes and nurse narrative notes can be considered as a promising data to facilitate ADR analysis. The tremendous values of the invisible ADR information can feasible derive from this underlying data, which are

Table 2.1: The characteristic of data sources in biomedical domain.

Data source	Data type	Noise-prone	Reliability	Data characteristics	Example resources
EMR	S, U	Medium	High	Patients demography, Discharge summary, Laboratory results, Longitudinal data	MIMIC <sup>39</sup> , OMOP <sup>1</sup> , I2B2 <sup>2</sup>
SRS	S, U	Medium	Medium	Volunteer report	FAERS <sup>3</sup>
Omics database	S, U	Low	High	Knowledge base	ChEMBL <sup>4</sup> , KEGG <sup>5</sup> , DrugBank <sup>6</sup> , STRING <sup>7</sup>
Biomedical literature	U	Low	High	Biomedical publication	MEDLINE/PubMed data <sup>8</sup>
Social media	U	High	Low	User post, comment, opinion in forum, blog	DailyStrength <sup>9</sup> , WebMD <sup>10</sup> , Twitter

**Data type:** S—Structured data, U—Unstructured data

<sup>1</sup>Observational Medical Outcomes Partnership —<http://omop.org>

<sup>2</sup>Informatics for Integrating Biology & the Bedside —<https://www.i2b2.org>

<sup>3</sup>FDA Adverses Events Reporting System —<https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm>

<sup>4</sup>Chemical database of bioactive molecules —<https://www.ebi.ac.uk/chembl>

<sup>5</sup>Kyoto Encyclopedia of Genes and Genomes —<http://www.genome.jp/kegg>

<sup>6</sup>A Pharmaceutical Knowledge Base —<https://www.drugbank.ca>

<sup>7</sup>Protein-Protein Interaction Networks —<https://string-db.org>

<sup>8</sup>Biomedical literature —[https://www.nlm.nih.gov/databases/download/pubmed\\_medline.html](https://www.nlm.nih.gov/databases/download/pubmed_medline.html)

<sup>9</sup>Heath forum —<https://www.dailystrength.org>

<sup>10</sup>Health information services website —<http://www.webmd.com>

proven by many studies of research. Nursing document or nursing narrative facilitates a real-time or synchronous patient's status as a recording of a timely log and a summary at the end of a shift. The deliverable messages are including the observable patient's health situation, assessment, plan, and recommendation to a next shift. Even though the nursing narrative contains enormous redundant data, but there are major advantages for patient monitoring and harmful changed status detection of a given certain condition, e.g., given the changing a dose of medicine and observe the patient's response. On the other hand, a discharge summary is a primary deliverable document to support communication among health professional teams in the hospital [43]. The content is recorded as a free text that summarizes a patient's hospitalization. Apart from the current admission information, significant finding, procedures and treatment, prescription medication, laboratory test and a result, it also conveys family history, illness history, and the follow-up instruction. Unlike nursing document, the discharge summary mostly captures the non-redundant and significant data instead of log data. The utilization is found in many pieces of research [44, 45, 46, 47, 48, 49] by deploying NLP technique to explore the potential ADR from this type of EMR document. Another dominant note, radiological report, contains a radiology imaging that is derived from an advanced imaging technology, and further free text data consolidation. A diagnostic radiologist, who specializes in the interpretation of these images, can take advantage of radiology imaging for diagnostic and disease treatment. The remaining free text in the report narrates the reason of examination, the underlying medical condition including the summarization of radiology examination and interpretation as a final report. This beneficial interpretation of radiology and patients condition information can contribute towards the ADR signal detection as well.

While EMR data is considered as the highly reliable source, the written-style by means of a medical professional is slightly painful from abbreviation, punctuation mark, typo, etc. Typically, a medical professional narrates patient condition and health status as a real-time therefore obtained textual data has the tendency to be no grammar strictly. However, EMR data is recognized as the promising source of healthcare data analytics because EMR data is plenty of rich implicit knowledge that is needed text mining for explicit knowledge extraction. The example of heterogeneous data in EMR

is depicted in Figure 2.3, and the promising sources of EMR data for clinical text mining are shown in Table 2.2. The following is the comparison between advantages and disadvantages of EMR data.

### Advantages

- High-reliability repository (data is derived by domain experts)
- Longitudinal patient care and outcome
- High clinical sensibility
- Structured and unstructured data

### Disadvantages

- Abbreviation, punctuation mark, moderately ill-format
- Non-publicly available, Data privacy

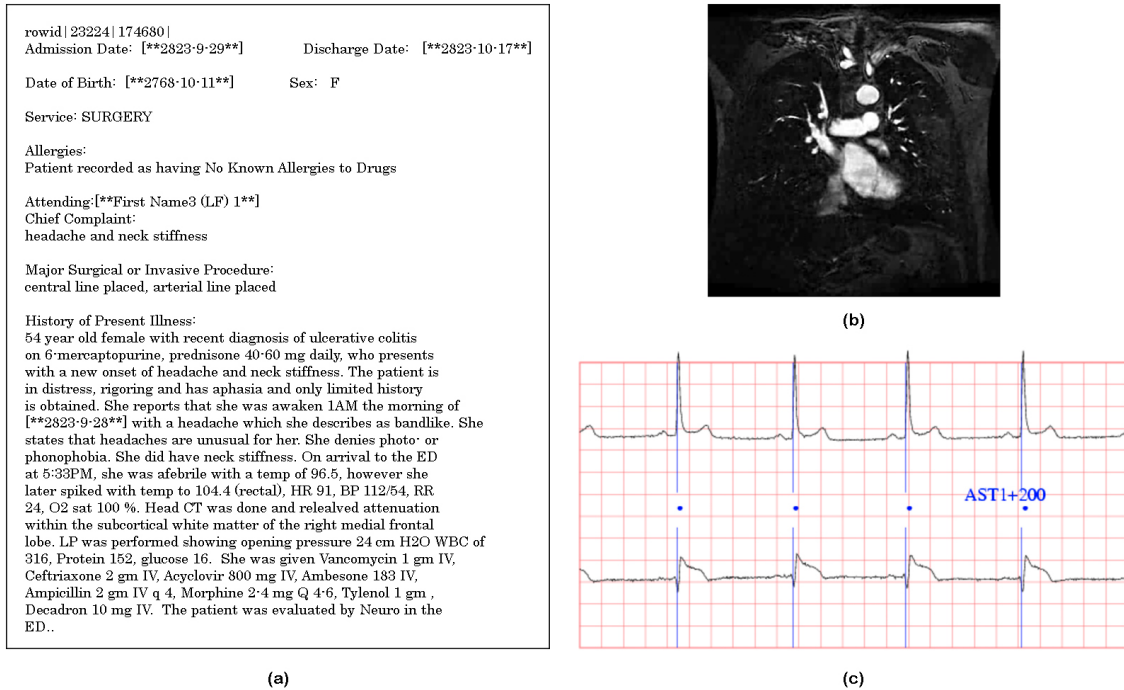


Fig. 2.3: Example of heterogeneous data in EMR. (a) Textual data in a discharge summary. (b) Magnetic Resonance Angiography (MRA) images. (c) ECG signal.

Table 2.2: The electronic medical record data sources for medical text mining.

Reference	Data source	Public	Description	Data Characteristic
[50]	CCAE <sup>1</sup>	N	MarketScan Commercial Claims and En-counters (USA) provide information on pharmacoepidemiologic data sources for use in epidemiology, health services research, healthcare economics.	Claim data, symptom and diagnosis data.
[50]	GE EHR <sup>2</sup>	N	GE Healthcare MQIC (Medical Quality Improvement Consortium) database.	A longitudinal outpatient population, and captures events in structured form that occur in usual care, including patient problem lists, prescriptions of medications, and other clinical observations as experienced in the ambulatory care setting.
[51, 44]	I2B2 <sup>3</sup>	Y	Informatics for Integrating Biology and the Bedside (i2b2).	The files contain a random selection of 100,000 records for each of 97 common lab tests, for a total of 9.7 million records.

*Continued on next page*<sup>1</sup><http://www.bridgetodata.org/node/987><sup>2</sup>[http://www.emr.msu.edu/Documents/mqic\\_main.htm](http://www.emr.msu.edu/Documents/mqic_main.htm)<sup>3</sup><https://i2b2.org/>

Table 2.2 – *Continued from previous page*

Reference	Data source	Public	Description	Data Characteristic
[52]	MIMIC	Y	Multiparameter Intelligent Monitoring in Intensive Care, Intensive Care Unit (ICU) patients.	Structured data-medical, surgical, coronary care, neonatal, laboratory test, disease diagnosis, etc.; unstructured clinical narrativesmedical note, nurse note, discharge summary; waveform data.
[45, 46, 47, 48, 50]	NYPH	N	New York Presbyterian Hospital at Columbia University Medical Center.	Containing of 1.2 million narrative notes; discharge summaries, operative reports, and reports from numerous ancillary services (e.g., radiology and pathology).
[53, 50]	OMOP <sup>4</sup>	N	Observational Medical Outcomes Partnership; An observational healthcare databases simulated data for studying the effects of medical products [47].	10 million persons; 90 million drug exposures; 5000 different drugs; 300 million condition occurrences; 4500 different conditions; over a span of 10 years; only 1.8% of the 20 million possible drug-condition combinations (population statistic from [53]).

*Continued on next page*<sup>4</sup><http://omop.org>

Table 2.2 – *Continued from previous page*

Reference	Data source	Public	Description	Data Characteristic
[54, 55]	STRIDE	N	The Stanford Clinical Data Warehouse.	Containing of 1.6 million patients; 15 million encounters; 25 million coded ICD-9 diagnoses, and a combination of pathology, radiology, and transcription reports; over 9.5 million unstructured clinical notes over a period of 17 years (population statistic from [55])
[56, 57]	The Stockholm EPR Corpus [58]	N	The electronic patients record from Karolinska University Hospital	Over 512 clinical units; over 2 m patients; structured data-age, gender, ICD-10 diagnosis code, drugs, laboratory result, admission and discharge time; unstructured data-clinical narratives.
[42]	VUMC	N	The Vanderbilt University Medical Center.	Inpatient and outpatient, clinical information, laboratory values, imaging and pathology reports, billing codes, and clinical narratives; 1.9 million patients with highly detailed longitudinal data for about 1 million.
[42, 59]	Korean tertiary teaching hospital clinical database	N	Korean tertiary teaching hospital clinical database.	32,033,710 prescriptions; 115,241,147 laboratory tests; 1,011,055 hospitalizations; 530,829 individual patients (Jan 2000 - Mar 2010).

## Spontaneous Report System (SRS)

Post-market surveillance is a necessary process to monitor the pharmaceutical drug safety after it has been released on the market. In 1964, the *Yellow card system* and the *Blue card system* are deployed as spontaneous reporting system in the United Kingdom and Australia respectively [60]. In the United State, *Med Watch*<sup>5</sup> has been used for the same purpose. The data from Med Watch is stored in Adverse Event Reporting System (AERS) database<sup>6</sup> and used for analysis by the US Food and Drug Administration (FDA).

A spontaneous reporting system is aimed to collect passive information relevant to adverse drug reaction that might be a cause of a severe adverse event, e.g., death, life-threatening, hospitalization, disability, congenital anomaly, etc. and identify existing unrecognized severe adverse reactions. The report is recorded by either professional clinicians who have suspected or diagnosed on a severe drug and clinical event concerns, or consumers and patients who have encountered an adverse event. Even though the report is provided by domain experts and experienced users, the report is often redundant, false reporting, or under-report. The redundant report might be found on the common adverse reaction that already notified in drug leaflets or package insert. While the inaccurate report relies on skill or experience of diagnosis or conclusion of a volunteer who provides information, it can lead to increasing of false alarm adverse reaction rate. On the one hand, the under-reporting is caused by some factors, e.g., the severity of the reaction, how is long that a drug has been on the market, whether an adverse reaction is already known by a reporter [61]. Such factors probably result to underestimating of the significance of a particular reaction [62]. Figure 2.4 depicts an example form of a spontaneous report in Med Watch.

### Advantages

- Population-based for rare event identification
- Report is based on clinical practice (not from clinical trials)
- Low cost data (collected by reporting)

---

<sup>5</sup><https://www.fda.gov/safety/medwatch/default.htm>

<sup>6</sup><https://www.fda.gov/drugs/informationondrugs/ucm135151.htm>

## Disadvantages

- Moderate reliability repository
- Under-reporting ADR
- Lacking in clinical sensibility
- Quality of provided information and missing data

## Omics Database

As large collections from various areas in biology, omics data provides a large pile of meaningful descriptions related to healthcare information. Some major data can be enumerated as follows. Proteomics can be defined as a complete set of proteins produced by a given cell or organism under a defined set of conditions [63], genomics is a studying about the genomes of organisms, e.g., genome sequencing, genome-wide association, gene expression analysis. UnitProt<sup>7</sup>, SwissProt<sup>8</sup>, PDB<sup>9</sup>, GO<sup>10</sup>, KEGG<sup>11</sup> are the sources of proteomics data. Metabolomics is the comprehensive, the qualitative, and the quantitative analysis of all the small molecules [64] or all chemical reactions involved in maintaining the living state of the cells and the organisms [65]. The related databases are such as ChEMBL<sup>12</sup>, PubChem<sup>13</sup>, DrugBank<sup>14</sup>, ChemSpider<sup>15</sup>, etc. Pharmacogenomics focuses on how human genes and complex gene systems influence the response to drugs [66]. The integration of KEGG, GO, DrugBank can provide information on this area. Toxicogenomics is a scientific field and an outgrowth of the human genome project. It is closely allied to pharmacogenetics that analyzes effects of heredity on responses of humans to drugs [67]. Furthermore, it describes the measurement of global gene expression changes in biological samples exposed to toxicants [68].

---

<sup>7</sup><http://www.uniprot.org>

<sup>8</sup>[http://web.expasy.org/docs/swiss-prot\\_guideline.html](http://web.expasy.org/docs/swiss-prot_guideline.html)

<sup>9</sup><https://www.rcsb.org>

<sup>10</sup><http://www.geneontology.org>

<sup>11</sup><http://www.genome.jp/kegg>

<sup>12</sup><https://www.ebi.ac.uk/chembl>

<sup>13</sup><https://pubchem.ncbi.nlm.nih.gov>

<sup>14</sup><https://www.drugbank.ca>

<sup>15</sup><http://www.chemspider.com>



Several databases of this type are KEGG, UniProt, SwissProt, STRING<sup>16</sup>, DrugBank, SIDER<sup>17</sup>. Figure 2.5 depicts an example data form DrugBank. The following is the comparison between advantages and disadvantages of omics data.

### **Advantages**

- High reliability repository
- Structured data (knowledge base in form of database)
- Many publicly available databases

### **Disadvantages**

- Lacking in clinical sensibility

## **Biomedical Literature**

Biomedical literature contains a scientific literature related to life science and medical domain. The contents in literature include knowledge, discovery, information, report, an experimental result related to the biomedical domain. Biomedical literature provides the principal advantage in a summary of document contents located at the abstract, and it can be found in every research publication. The abstract of any literature typically contains a key content in summary format that is easier for mining. The literature data source provides continuously updated information due to published new research articles every day. Moreover, a writing style is carefully and well written-form, fine controlled vocabulary and less syntactic or grammar error. The content in medical literature is based on experiment, discovering or evidence, therefore, it is a high reliability. However, the patient-based information is excluded from the text, and it is no narrative information relevant temporal events during patient admission such as health status of patient before or after taken a treatment etc. Therefore, the weakness of this data is unable to express viewpoint of patient dimension.

---

<sup>16</sup><https://stringdb.org>

<sup>17</sup><http://sideeffects.embl.de>

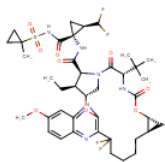
<b>Name</b>	<b>Voxilaprevir</b>
<b>Accession Number</b>	<b>DB12026</b>
<b>Type</b>	Small Molecule
<b>Groups</b>	Approved
<b>Description</b>	<p>Voxilaprevir is a Direct-Acting Antiviral (DAA) medication used as part of combination therapy to treat chronic Hepatitis C, an infectious liver disease caused by infection with Hepatitis C Virus (HCV). HCV is a single-stranded RNA virus that is categorized into nine distinct genotypes, with genotype 1 being the most common in the United States, and affecting 72% of all chronic HCV patients [3].</p> <p>Voxilaprevir exerts its antiviral action by reversibly binding and inhibiting the NS3/4A serine protease of Hepatitis C Virus (HCV) [FDA Label]. Following viral replication of HCV genetic material and translation into a single polypeptide, Nonstructural Protein 3 (NS3) and its activating cofactor Nonstructural Protein 4A (NS4A) are responsible for cleaving genetic material into the following structural and nonstructural proteins required for assembly into mature virus: NS3, NS4A, NS4B, NS5A, and NS5B [2]. By inhibiting viral protease NS3/4A, voxilaprevir therefore prevents viral replication and function. Treatment options for chronic Hepatitis C have advanced significantly since 2011, with the development of Direct Acting Antivirals (DAAs) such as voxilaprevir.</p>
<b>Structure</b>	 <div> <input type="text"/> <input type="button" value="Download"/> <input type="button" value="Similar Structures"/> </div>
<b>Synonyms</b>	Not Available
<b>External IDs</b>	GS 9857 / GS-9857 / GS9857
<b>Product Ingredients</b>	Not Available

Fig. 2.5: An example data from DrugBank.

One of large-scale biomedical literature repositories is MEDLINE<sup>18</sup>. It contains more than 24 million literatures since 1809 and spends hard disk space up to 16.9 GB. MEDLINE is publicly available at U.S. National Library of Medicine website<sup>19</sup>. The data is in xml format which is recognized as semi-structured data, however, the abstract itself is unstructured text. Figure 2.6 depicts an example of MEDLINE data. Using biomedical literature for the purpose of ADR extraction, researchers typically retrieve data by querying abstract or Medical Subject Headings (MeSH) and generate hypotheses concerning specific drugs and health problems. Similarly, IEEE Xplore is a

<sup>18</sup><https://www.ncbi.nlm.nih.gov/pubmed>

<sup>19</sup>[https://www.nlm.nih.gov/databases/download/pubmed\\_medline.html](https://www.nlm.nih.gov/databases/download/pubmed_medline.html)

digital library of a high-quality literature repository in engineering and technology including biomedical research area. It provides application programming interface (API) named as IEEE Xplore Search Gateway<sup>20</sup> for querying abstract of an article that is published with Institute of Electrical and Electronics Engineers. Figure 2.7 depicts an example of IEEE Xplore API data. Others API for publicly available literature repository are such as arXiv API<sup>21</sup>, Elsevier Scopus APIs<sup>22</sup>, National Library of Medicine (NLM) API<sup>23</sup>. The following is the comparison between advantages and disadvantages of biomedical literature data.

### **Advantages**

- High reliability repository (data is derived by experiment-based results)
- Well written-form
- Low cost data
- Publicly available data

### **Disadvantages**

- Lacking in clinical sensibility
- Lacking in temporal event
- Falling into biased data due to passive outcome monitoring

## **Social Media**

The social network and social media become a famous source of data nowadays, due to the influence of Web 2.0 technology. Many healthcare communities emerge in social networks. Such variety medical forums can provide, service or exchange healthcare knowledge among users, e.g., DataGeno, PatientsLikeMe, and WebMD. The growth of online social networking forums brings to patients voluntarily sharing their experience

---

<sup>20</sup><http://ieeexplore.ieee.org/gateway>

<sup>21</sup><https://arxiv.org/help/api/index>

<sup>22</sup>[https://dev.elsevier.com/sc\\_apis.html](https://dev.elsevier.com/sc_apis.html)

<sup>23</sup>[https://wwwcf.nlm.nih.gov.nlm\\_eresources/eresources/search\\_database.cfm](https://wwwcf.nlm.nih.gov.nlm_eresources/eresources/search_database.cfm)

```

<Article PubModel="Print">
  <Journal>
    <ISSN IssnType="Print">0021-8820</ISSN>
    <JournalIssue CitedMedium="Print">
      <Volume>40</Volume>
      <Issue>1</Issue>
      <PubDate>
        <Year>1987</Year>
        <Month>Jan</Month>
      </PubDate>
    </JournalIssue>
    <Title>The Journal of antibiotics</Title>
    <ISOAbbreviation>J. Antibiot.</ISOAbbreviation>
  </Journal>
  <ArticleTitle>Semisynthetic beta-lactam antibiotics. III. Effect on antibacterial activity and comt-susceptibility of chlorine-introduction into the catechol nucleus of 6-[(R)-2-[3-(3,4-dihydroxybenzoyl)-3-(3-hydroxypropyl)-1-ureido]-2-phenylacetamido]penicillanic acid.</ArticleTitle>
  <PageRange>
    <MedlinePgn>22-8</MedlinePgn>
  </PageRange>
  <Abstract>
    <AbstractText>The resistance of 6-[(R)-2-[3-(3,4-dihydroxybenzoyl)-3-(3-hydroxypropyl)-1-ureido]-2-phenylacetamido]penicillanic acid (1a) to metabolism by catechol-O-methyl-transferase (COMT) was increased by introduction of the chlorine atom into the catechol moiety. Penicillins (1b-1d) having one or two chlorine atoms at the positions adjacent to the hydroxyl group were found to have greater stability to COMT. This resulted in greater efficiency in vivo in experimental Pseudomonas aeruginosa and Escherichia coli infections. In vitro activities were essentially unchanged.</AbstractText>
  </Abstract>
  <AuthorList CompleteYN="Y">
    <Author ValidYN="Y">
      <LastName>Ohi</LastName>
      <ForeName>N</ForeName>
      <Initials>N</Initials>
    </Author>
  </AuthorList>
  ■ ■ ■
  <MeshHeadingList>
    <MeshHeading>
      <DescriptorName UI="D000818" MajorTopicYN="N">Animals</DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName UI="D000900" MajorTopicYN="N">Anti-Bacterial Agents</DescriptorName>
      <QualifierName UI="Q000138" MajorTopicYN="Y">chemical synthesis</QualifierName>
      <QualifierName UI="Q000494" MajorTopicYN="N">pharmacology</QualifierName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName UI="D001419" MajorTopicYN="N">Bacteria</DescriptorName>
      <QualifierName UI="Q000187" MajorTopicYN="N">drug effects</QualifierName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName UI="D065098" MajorTopicYN="Y">Catechol O-Methyltransferase Inhibitors</DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName UI="D004927" MajorTopicYN="N">Escherichia coli Infections</DescriptorName>
      <QualifierName UI="Q000188" MajorTopicYN="N">drug therapy</QualifierName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName UI="D007202" MajorTopicYN="N">Indicators and Reagents</DescriptorName>
    </MeshHeading>
  </MeshHeadingList>

```

Fig. 2.6: An example of MEDLINE data in XML format.

on drug use, making these forums as valuable resources for ADR analysis [69, 70, 71]. Recently, the information extraction from the healthcare forum, Medications.com, has been proposed [72]. The source of data is the precious resource of sharing experience

```

▼<root>
  <totalfound>59</totalfound>
  <totalsearched>4302190</totalsearched>
  ▼<document>
    <rank>11</rank>
    ▼<title>
      ▼<![CDATA[
        Characterization of CuInSe<inf>2</inf> thin films prepared by ion-beam sputtering
      ]]>
    </title>
    ▼<authors>
      ▼<![CDATA[
        Zhuang-hao Zheng; Ping Fan; Dong-ping Zhang; Xing-Min Cai; Guang-xing Liang
      ]]>
    </authors>
    ▼<affiliations>
      ▼<![CDATA[
        College of Physical Science and Technology, Institute of Thin Film Physics and Applications,
        Shenzhen University, Shenzhen, 518060, China
      ]]>
    </affiliations>
    ▼<controlledterms>
      ▼<term>
        <![CDATA[ annealing ]]>
      </term>
      ▼<term>
        <![CDATA[ copper compounds ]]>
      </term>
      ▼<term>
        <![CDATA[ indium compounds ]]>
      </term>
      ■ ■ ■
    </controlledterms>
    ▼<pubtype>
      <![CDATA[ Conference Publications ]]>
    </pubtype>
    ▼<publisher>
      <![CDATA[ IEEE ]]>
    </publisher>
    ▼<py>
      <![CDATA[ 2009 ]]>
    </py>
    ▼<spage>
      <![CDATA[ 1 ]]>
    </spage>
    ▼<epage>
      <![CDATA[ 2 ]]>
    </epage>
    ▼<abstract>
      ▼<![CDATA[
        CuInSe<sub>2</sub> (CIS) thin films were prepared by ion-beam sputtering at different
        substrate temperatures. The films prepared at room temperature were annealed at different
        temperatures. Films annealed at appropriate temperatures are dense, uniform and of single-
        phase.
      ]]>
    </abstract>
    ▼<isbn>
      <![CDATA[ New-2005_POD_978-1-4244-3829-7 ]]>
    </isbn>
  </document>

```

Fig. 2.7: An example of IEEE Xplore API data in XML format.

about drugs use among voluntary patients. Firstly, the implementation of Information Retrieval module can retrieve messages posted from the crawl technology. Secondly,

applying text processing module can obtain an appropriate data format. Thirdly, information extraction is manipulated to perform the named entity recognition using drug name and adverse reaction as a dictionary. Finally, hidden markov model is carried out to find the relationship between drugs and their adverse reactions in relationship extraction module.

From the five sources of healthcare data mentioned in Table 2.1, the social network data provides the lowest reliability and contains vast of noise-prone. In particular, data from forum, blog or twitter, is usually derived from non-domain expert users, in other words, by patients or relative of patients through their opinions or comments that are based on emotional or sentimental rather than factual data expression. Nevertheless, due to large-scale data, many research works [73, 74, 75] have been explored such source of data on healthcare domain. Figure 2.8 depicts an example of tweets data. The following is the comparison between advantages and disadvantages of social network and social media data.

### **Advantages**

- Population-based for rare event identification
- Low cost data
- Publicly available data

### **Disadvantages**

- Low reliability repository (data is derived by non-domain experts)
- Comments or opinion based on emotional or sentimental data expression
- Poor text quality , ill-format, noise-prone
- Lacking in clinical sensibility

## **2.3 Named Entity Recognition**

Normally, a named entity is an individual word or a set of words that indicates a real-world object, e.g., *Shinzo Abe*, *Tokyo*, *Rakuten*. Slightly different to a named entity



Fig. 2.8: An example of Tweets data.

in biomedical text, a real-world object is relevant to a medical object, e.g., *Cyclobenzaprine*, *Gastroesophageal Reflux*, *Tracheal tube*. The named entity recognition is a process involving the identification of named entities in text and classification them into a set of predefined categories, e.g., *person*, *organization*, *location* for a general domain or *drug name*, *disease name*, *device* for the medical domain. Given the following a fragment of a sentence as an example,

*“The International Olympic Committee awarded the Games of the XXXII Olympiad in 2020 to Tokyo.”*

the extracted named entities of three categories; organization, location and date, are the following information,

Organization: *International Olympic Committee*,

Location: *Tokyo*,

Date: *2020*.

The extracted information can provide more comprehension corresponding domain of a given text and can be used in other processes of text mining such as relation extraction, information retrieval.

The medical named entity recognition is a fundamental and an essential process towards efficiency for text mining. The spelling correction, medical text mention recognition, and normalization are typically involved in this process. In biomedical text mining from EMR, named entity recognition task aims to provide a precise system to recognize medical attributes concerning medications, adverse reactions, anatomical parts, diseases, devices, temporal phrases, etc. For example in the following sentence,

*“Patient was found to have mild thrush, treated with Nystatin, viscous lidocaine.”*

the extracted named entities are the following information,

Drug name: *Nystatin* and *Viscous lidocaine*,

Disease name: *Thrush*.

The normalization task unifies such medical attributes into common lexicons based on the identical semantic, so-called *concept*. The Concept Unique Identifier<sup>24</sup> (CUI) defined by Unified Medical Language System (UMLS) is widely utilized for such purpose and be used as a referral identifier. According to the above example regarding biomedical text, drug name “*Nystatin*” and “*Viscous lidocaine*” can be normalized to concept CUI *C0028741* and CUI *C0721362* respectively, and “*Thrush*” can be normalized to concept CUI *C0006840*. Named entity recognition challenges in biomedical text analysis can be found in several shared tasks such as SemEval [76, 77, 78], ShARe/CLEF eHealth [52], i2b2/VA challenge [79], GermEval [80], etc.

---

<sup>24</sup>a concept in the UMLS Metathesaurus

One of notable biomedical named entity recognition tools is MetaMap that is developed by Aronson [81]. The tool relies on natural language processing technique to recognize biomedical text to formal UMLS Metathesaurus. MetaMap is widely used in literature related to biomedical text mining and publicly accessible<sup>25</sup>. The National Library of Medicine provides MetaMap service usage via the online version as interactive, batch, and Web API modes and the off-line version of Java API. The algorithm inside MetaMap uses a knowledge-intensive approach by handling natural language processing tasks such as tokenization, part-of-speech tagging, the lexical lookup from knowledge based, syntactic analysis, negation extraction, word sense disambiguation [82]. MetaMap provides not only UMLS concept mapping but also UMLS semantic network that consists of a set of hierarchy information of two levels categorization; semantic group [83][84] and semantic type [85]. The 15 semantic groups and 133 semantic types are available for UMLS concept mapping. The biomedical named entity recognition, sentence boundary detection and normalization will be discussed in data preparation section of the next chapter (see Section 3.2).

## 2.4 Open Information Extraction

While the traditional information extraction entirely requires precisely target relation beforehand, Open Information Extraction (OpenIE) aims to extract relations without predefined relation category and independent domain. The OpenIE paradigm is well-known as a generalization of conventional information extraction that can potentially deal with large-scale corpora without manual tagging of relations [86].

Early of OpenIE [87] aims to extract an unknown relation in advance on highly scalable web corpus. The evident achievements on web mining deliver to an extensive paradigm shift in biomedical text mining. Currently, there are many numbers of effective information extraction methods. The REVERB [88], the double precision of the traditional OpenIE (TEXTRUNNER [89, 90]), retrieves the verb-based relation phrases through two constraints corresponding syntactic constraint and lexical constraint. The relation phrases are derived by shallow parsing, and the both constraints are applied to

---

<sup>25</sup><https://metamap.nlm.nih.gov>

overwhelm two significant of incoherent and uninformative extraction problems. The former problem is related to no meaningful interpretation of extracted relation phrase, and the later one is correlated with omitting the critical information. The OLLIE [91], the next revision of OpenIE, is introduced by the same group of researchers with REVERB that aims to overcome the weaknesses of verb-based reconciled relation phrase and context disregard. In their work, a bootstrapping method is used to create a large corpus of open pattern templates using seed tuples derived by REVERB. In the extraction process, the pattern matching based on predefined templates is extended with nouns, adjectives, etc. Additionally, the discovered relations are considered on the context relevance for hypothetical or conditionally true expansion. Recently, the Stanford Natural Language Processing group develops OpenIE tool [92] to reduce a large pattern set of canonical sentences and excerpt self-contained clauses from longer sentences as well.

## 2.5 Distant Supervision

Relation extraction with supervised learning is proven to be successful in many previous studies [93, 94, 95], however, the achievement of supervised learning method is limited to the amount of training set [96]. The training data acquisition requires human efforts for manual label tagging that is tedious and laboring task. An alternative solution is utilized knowledge base for data curation by distant supervision method.

The distant supervision aims to overcome the fundamental limitations of unavailable labeled examples that are usually done by human labor. This paradigm has a comparative advantage over manual data tagging regarding inexpensive, less time-consuming and feasible for large-scale corpora. The early work on distant supervision is achieved by Craven, M. et al. [14]. In their work, a term *weakly labeled data* is presented for biomedical relation extraction from MEDLINE. Lately, Mintz, M. et al. [15] proposes an interchangeable paradigm, *distant supervision*, to extract relation from Freebase. Their assumption relies on “*if the two entities participate in a relation, any sentence that contains those two entities might express that relation.*”.

In medical text mining, particularly ADR identification problem, there is a few

work that leverages distant supervision paradigm. The work of Segura-Bedmar, I., et al. [97], the Spanish social media from health forum is collected to explore plausible adverse reaction. The shallow linguistic is examined on a sentence of any pair of drug and clinical event that is appeared to co-occur within a window size of 250 tokens.

The similar manner as the work of Segura-Bedmar, I., et al., a label of each pair of drug and event, that is found to appear together in the same sentence, is derived by projection from the knowledge bases, namely *entity-level*. While a label for each sentence instance that appears a same of a drug-event pair, known as *instance-level*, cannot be acquired by such projection method. Therefore, the consideration of the two levels labeling is needed for model classification, and it needs a particular method such as multiple-instance learning (MIL) to deal with such situation. The followings example describes the concept of entity- and instance-levels and their problem as mentioned above. Considering the followings two sentences,

*“He was put on a **dopamine**, though this was discontinued due to persistent **tachycardia**.”*,

*“Medication history of the patient is a **dopamine** and a dobutamine, he was checked by EKG for **tachycardia** before admission.”*

Both sentences contain a drug “dopamine” and a clinical event “tachycardia”. Suppose that a pair of dopamine and tachycardia exists in knowledge base from SIDER, which is the database of adverse drug reaction (ADR) relation. The goal of distant supervision is to look up a pair of  $\text{dopamin}_{drug}$  and  $\text{tachycardia}_{event}$  in knowledge base SIDER and infer relation label. From the example, it is clearly shown that the entity-level of a drug  $\text{dopamin}_{drug}$  and an event  $\text{tachycardia}_{event}$  expresses true relation as inferred from SIDER. However, the instance-level is expressed as true relation corresponding ADR only in the first sentence, but the second sentence is not expressed as the true relation. This problem is well known as the noisy label.

Recently, there are attempt to employ distant supervision as data labeling for ADR identification [97, 98] or integrate with a human label to extensive training size [99]. The employing of distant supervision for relation extraction can be divided into two

main approaches.

- The first approach makes use knowledge base that contains entity pairs of interest as a guideline to extract patterns from such entity pairs. All extracted patterns are assumed to be relevant to relation label and can be used for discovering unknown entity pairs. This approach can be thought as feature labeling that considers on an association between pattern and relation label, which is similar to the pattern bootstrapping method. The main challenge of this approach is how to filter noisy patterns out of qualified patterns. The utilization of distant supervision as a guideline for feature labeling in bootstrapping method will be discussed later in Chapter 4.
- The latter approach utilizes facts from knowledge base to construct training dataset instead of manual label tagging. This approach can be seen as instance labeling and much more complicated than the feature labeling. The mapping entity pairs from knowledge base to text in a sentence in order to infer relation label, it usually contains noise because some labeled sentence might not express true relation. Therefore as mentioned above, a particular feature engineering approach or model classifier is needed for such problem. The utilization of distant supervision as a set of training examples for biomedical relation extraction problem will be discussed later in Chapter 5.

# Chapter 3

## Relation Extraction in Clinical Text

### 3.1 Drug-Event Relation Extraction

The drug-event relation extraction from unstructured text can be defined as the method to discover hidden drug relevant clinical event from a given text that might express its relation as harmful or beneficial outcomes from drug administration. While harmful or unintended symptom caused by a drug is known as an adverse drug reaction, the beneficial result from treatment by medicine is the so-called therapeutic indication. With significant increasing of toxicology and clinical safety failures, drug safety and post-marketing surveillance have become a crucial topic. The World Health Organization gives some key facts that motivate ADRs research as follows: ADRs are among leading causes of deaths and effect to in every country, the majority of ADRs are preventable, ADRs probably related to costs and no medicine is risk-free [100].

In a computational method for drug-event relation extraction, there are multidisciplinary techniques along the variety of data types and the objectives. The following is the summary of a group of methods for drug-event pair relation identification and Table 3.1 exhibits the details of each method.

- **Co-occurring method** An evidence-based method is inspected for associative co-occurring between drug and event, the chi-square ( $\chi^2$ ) statistic and its  $p$ -value are used to test the hypothesis of no association [45, 46, 47, 101, 102].
- **Ranking method** Relation candidate generation is generated by disproportional

analysis of co-occurring drug-event pair [53, 55, 103] or by classifier model [53], then rank its outcomes.

- **Rule-based method** In this categorization, rule-based and pattern based are grouped together. Rule based is used as a template for rule matching to discover hidden drug-event pair relation. Table mapping [104], part of speech tagging [105], association rule mining [106] or regular expression [107] can be used for rule (or pattern) generation.
- **Machine learning** The method mainly focuses on deploying of feature engineering and uses machine learning method such as SVM, Naïve Bayes, Decision Tree for relation classification [50, 108, 109, 110].
- **Distant supervision** The method makes use of knowledge base for training data labeling instead of manual label tagging [97, 99] (see Chapter 4 and Chapter 5 for more details).

Table 3.1: A list of previous studies on drug-event relation extraction

Reference	Problem	Material and Study Population	Method
Chen E.S. et al. [45]; Co-occurring method	Identify association between drug-disease pair	<ul style="list-style-type: none"> <li>– Discharge summary in the 2003-2004 (48,360 reports).</li> <li>– MDELIN article in the 2006 (81,828 related articles).</li> <li>– To investigate all drugs related to 8 diseases of interest</li> <li>– NLP tool; BioMedLee, MedLEE.</li> </ul>	<ul style="list-style-type: none"> <li>– Drug-disease annotation to highest-level MeSH descriptor and UMLS concept by BioMedLee and MedLEE tools.</li> <li>– Association examination of drug-disease pairs by co-occurrence and <math>\chi^2</math>.</li> <li>– Comparing the association derived across different annotation methods and data sources to evaluate the overall agreement.</li> <li>– Manual review process by a medical expert.</li> </ul>
Wang X. et al. [46]; Co-occurring method	To characterize phenotypic and environmental associations obtained from clinical reports	<ul style="list-style-type: none"> <li>– Discharge summary from NYPH (25,074 reports) in the 2014.</li> <li>– 1,997 unique drug concepts in scope.</li> <li>– 732 unique symptom concepts in scope.</li> <li>– 947 unique disease concepts in scope</li> <li>– NLP tool; MedLEE.</li> </ul>	<ul style="list-style-type: none"> <li>– Drug-disease annotation to UMLS concept.</li> <li>– Evaluation the association between disease-disease, drug-disease/symptom and disease-symptom by performing hypothesis testing <math>\chi^2</math>.</li> <li>– Deriving an indirect drug-adverse reaction by the computation of mutual information</li> </ul>
Wang X. et al. [47]; Co-occurring method	Detect associations between drug and adverse reaction	<ul style="list-style-type: none"> <li>– Discharge summary from NYPH in the 2004.</li> </ul>	<ul style="list-style-type: none"> <li>– Annotating drug-disease to UMLS concept.</li> <li>– Filtering confounding factors such as diseases occurring before the drug usage, etc.</li> <li>– Determining drug-adverse event co-occurring pairs and hypothesis testing based on <math>\chi^2</math>.</li> </ul>

*Continued on next page*

Table 3.1 – *Continued from previous page*

Reference	Problem	Material and Study Population	Method
Liu M. et al. [101]; Co-occurring method	Detect drug and adverse reaction association	<ul style="list-style-type: none"> <li>– Dataset of Yoon et al. [120] 470 drug-event pairs</li> <li>– Dataset from VUMCs EMR 187,595 patients record with 378 drugevent pairs</li> </ul>	<ul style="list-style-type: none"> <li>– Separating retrospective observations into two groups; study group and comparison group, based on rule based.</li> <li>– Determining drug-adverse event by hypothesis testing base on <math>\chi^2</math>, PRR, ROR, Yules Q (YULE), BCPNN, and GPS.</li> <li>– Evaluation matrices for performance assessment by precision, recall, and F-score.</li> </ul>
Roitmann E. et al. [102]; Co-occurring method	Clustering patients based on drug-event profile.	<ul style="list-style-type: none"> <li>– 6,011 patient records from clinical narratives from Danish mental health center in the 1998 to 2010.</li> <li>– Drug information (ATC), drug dosages, prescription intervals, and diagnosis code (ICD10).</li> </ul>	<ul style="list-style-type: none"> <li>– Identifying 2,347 patients with history of at least one drug and one adverse event.</li> <li>– Constructing patient vectors with 1,190 adverse event dimensions by tf-idf weighted values.</li> <li>– Stratifying the patients based on cosine dissimilarity profile.</li> <li>– Computing co-occurring score and weighted edges to analyze the cluster adverse event.</li> </ul>
Harpaz R. et al. [103]; Ranking method	ADR signal detection	<ul style="list-style-type: none"> <li>– Clinical narrative text from NYPH in the 2004-2010 and Adverse event report system (AERS) of the Food and Drug Administration in the 1968-2101Q3.</li> <li>– NLP tool-MedLEE.</li> <li>– RxNorm, UMLS concept (2011AA), MedDRA(V.13.1).</li> </ul>	<ul style="list-style-type: none"> <li>– Annotating the clinical narratives to UMLS concept.</li> <li>– Detecting drug-event association by disproportionality analysis and ranking.</li> </ul>

*Continued on next page*

Table 3.1 – *Continued from previous page*

Reference	Problem	Material and Study Population	Method
Duan L. et al. [53]; Ranking method	Rare ADR detection	– Simulated OMOP dataset.	<ul style="list-style-type: none"> <li>– Deploying the ensemble methods for drug-event detection; 2x2 contingency table, likelihood ratio and a Bayesian network.</li> <li>– Computing scores on probability outcomes from such three models and ranking.</li> </ul>
LePend P. et al. [55]; Ranking method	Analyzing patterns of off-label drug usage	<ul style="list-style-type: none"> <li>– Clinical narratives from STRIDE.</li> <li>– Drug indication data from Medi-Span for evaluation</li> <li>– NLP Tools-NCBO annotator, NegEx trigger</li> <li>– NCBO BioPortal library, RxNorm, SNOMED CT.</li> </ul>	<ul style="list-style-type: none"> <li>– Annotating terms in clinical notes using NCBO Annotator.</li> <li>– Applying NegEx trigger rules to separate negated terms and term normalization.</li> <li>– Constructing bag-of-terms.</li> <li>– Creating drug-indication associations using sliding window.</li> <li>– Filtering confounding factors.</li> <li>– Scoring the association by ROR and ranking.</li> </ul>
Park M. Y. et al. [104]; Rule-based method	Detecting the signals of ADR focused on laboratory abnormalities after treatment with medication	<ul style="list-style-type: none"> <li>– EMR data from Ajou University Hospital, in Korea from Jan 2000 - Mar 2010.</li> <li>– Laboratory anomaly is determined as adverse reaction.</li> <li>– 56 ADEs of interest from UpToDate Drug Information Database.</li> </ul>	<ul style="list-style-type: none"> <li>– Retrieving the list of known ADEs related to the selected drugs.</li> <li>– Constructing the mapping table to link between laboratory abnormalities detected by CERT algorithm and each known ADEs.</li> </ul>

*Continued on next page*

Table 3.1 – *Continued from previous page*

Reference	Problem	Material and Study Population	Method
Skentzos S. et al. [105]; Rule-based method	Identify ADR to Statins	<ul style="list-style-type: none"> <li>– Clinical narratives of outpatients from Partners Enterprise Allergy Repository (PEAR) in the 2000-2010.</li> <li>– 3,175 narrative notes.</li> <li>– UMLS concept, MedDRA code.</li> </ul>	<ul style="list-style-type: none"> <li>– Annotating medical terms to UMLS concept and MedDRA code.</li> <li>– Deploying parse tree and manual word class with semantic customization</li> </ul>
Ji Y. et al. [106]; Rule-based method	Identify causal relationships between drugs and their associated adverse drug reactions (ADRs)	<ul style="list-style-type: none"> <li>– Clinical narratives from the Veterans Affairs Medical Center in Detroit, Michigan.</li> <li>– 1,021 patients related to drug of interest 1,290 ICD9 codes associated with drug of interest.</li> </ul>	<ul style="list-style-type: none"> <li>– Detecting the relation candidate between drug-event pair by co-occurring.</li> <li>– Examining the association using experience-based fuzzy RPD model and the exclusive causal relation <math>\text{supp}(X \rightarrow Y)</math>.</li> </ul>
Iqbal E. et al. [107]; Rule-based method	Identify instance of adverse drug events (ADEs)	<ul style="list-style-type: none"> <li>– Clinical narratives from Clinical Record Interactive Search (CRIS) system, the South London and Maudsley NHS Foundation Trust (SLaM) (17,995 patients) in the 2007-2013.</li> <li>– NLP tools; GATE [111], Java Annotation Patterns Engine (JAPE).</li> </ul>	<ul style="list-style-type: none"> <li>– Constructing clinical event dictionary, including synonym and alternative spelling</li> <li>– Extracting clinical event from narrative text by GATE.</li> <li>– Rule generating by JAPE.</li> <li>– Performing bootstrapping method to iterative creating new and improve rule from misclassification.</li> </ul>

*Continued on next page*

Table 3.1 – *Continued from previous page*

Reference	Problem	Material and Study Population	Method
Li Y. et al. [50]; Feature-based method	ADR detection	<ul style="list-style-type: none"> <li>– EHR from NYP/CUMC and GE MQIC-including admission notes, discharge summaries, lab tests, structured diagnosis (ICD-9) codes and structured medication lists.</li> <li>– Claim data, CCAE.</li> <li>– Spontaneous reports from FAERS in the 2004-2010.</li> <li>– STITCH, MedDRA.</li> </ul>	<ul style="list-style-type: none"> <li>– Deploying LASSO to obtain the confounding adjusted signal score for each drug-event pair from NYP/CUMC and FAERS.</li> <li>– Normalizing ADR signal scores using p-value.</li> <li>– Combining scores of discovered drug-event pairs across different data sources.</li> </ul>
Peissig P. L. et al. [108]; Feature-based method	To classify patients risk	<ul style="list-style-type: none"> <li>– Healthcare data from CattailsMD EHR-Research Data Warehouse (RDW), Marshfield Clinic in the 1979-2011.</li> <li>– EHR data-diagnoses, procedures, laboratory results, observations, and medications for patients.</li> </ul>	<ul style="list-style-type: none"> <li>– Identifying training set of POS, NEG, and BP (borderline positive) samples.</li> <li>– Deploying Inductive Logic Programming (ILP) for rule learning as feature for classifier.</li> <li>– Constructing classifiers; Random forest, SMO, PART, J48, JRIP.</li> </ul>
Liu Y. et al. [109]; Feature-based method	Discriminating the drug-adverse event pairs from the drug-indication pairs	<ul style="list-style-type: none"> <li>– Narrative notes from STRIDE EMR database, over 9 million notes.</li> </ul>	<ul style="list-style-type: none"> <li>– Annotating textual medical records to UMLS concept.</li> <li>– Constructing drug-event association as a set of features by considering on patients timeline.</li> <li>– Building SVM classifier.</li> <li>– Evaluation method using 100-fold cross validation and independent validation set.</li> </ul>

*Continued on next page*

Table 3.1 – *Continued from previous page*

Reference	Problem	Material and Study Population	Method
Karlsson I. et al. [110]; Feature- based method	ADR prediction	– EMR from the Stockholm EPR Corpus in the 2009-2010.	<ul style="list-style-type: none"> <li>– Constructing feature vector corresponds to 1,312 drugs, 9,863 diagnosis code, age, and gender.</li> <li>– Modeling with two machine learning methods; Random forest and JRIP rule learner.</li> <li>– Evaluation using 10-fold cross validation</li> </ul>

## 3.2 Data Preparation

As the discussion in section 2.2, EMR is a rich source of individualized clinical data, which has a great potential for improving identification of patients experiencing adverse reactions, across drugs and indication areas. Narrative notes in EMR have been demonstrated as the promising source of the repository and widely utilized for such purpose [112, 113, 114, 115, 98]. In this dissertation, EMR from Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) database is used to analyze association between drug and clinical event whether adverse reaction or indication. The preprocessing is needed as a general medical text that composes of tokenization, stemming, named entity recognition and sentence boundary detection.

### 3.2.1 Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III)

MIMIC-III[116] is the notable publicly available source of EMR repository. The database is introduced by *National Institute of Biomedical Imaging and Bioengineering* and available at *PhysioNet*<sup>1</sup>. The over 58,000 hospital admissions for 38,645 adults and 7,7875 neonates are presented in the data source with spanning up to 12 years from June 2001. The rich information through the narrative texts is over 2 million event notes concerning 15 note categories.

Figure 3.1 depicts the ratio of total event notes per category and Table 3.2 described the purpose of the narrative text in each category. The highest portion is the nursing notes in which contain daily reports of patient condition and progress to be used for communicating the current health status. The second rank ratio is the radiology document that relevant to medical images or radioactive substance or sound wave such as X-ray, MRI, etc. However, the promising narrative text is a discharge summary that contains about 55,177 documents from 46,520 distinct patients. The discharge summary from EMR includes the long narrative texts with temporal information in various aspects. There are not only the summary of hospital course but also a history of patient illness, past medical history, medical on admission, discharge medications,

---

<sup>1</sup><https://mimic.physionet.org>

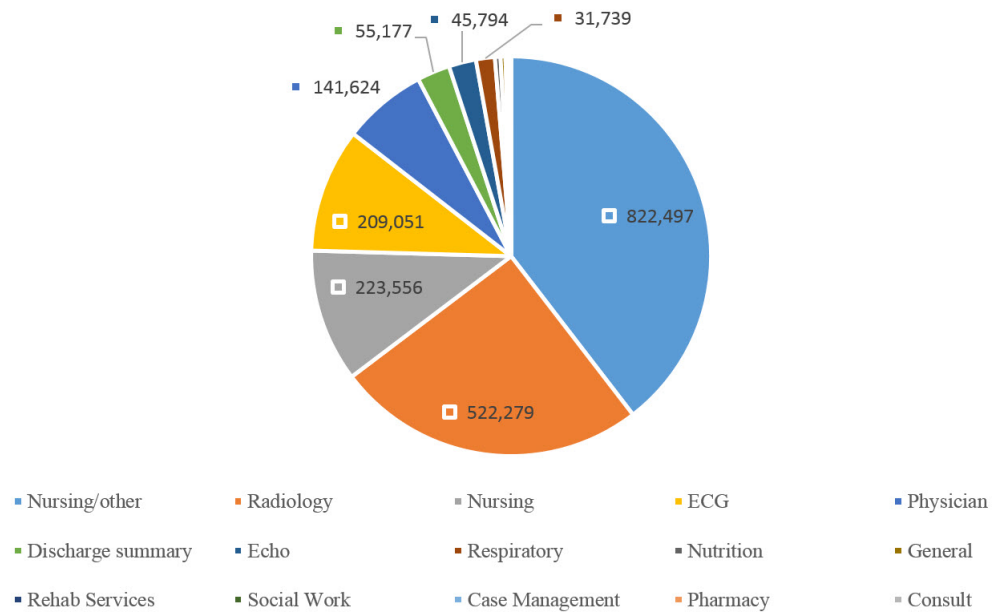


Fig. 3.1: MIMIC-III: a portion of event notes per category

physical exam, allergies information, discharge diagnosis etc.

Table 3.2: Statical number of note event categories from EMR and example of sentences.

Category	Size <sup>1</sup>	Example of Sentences
Nursing/other	822,497	RESP CARE: Pt remains intubated/on vent on CMV 500/14/.60 15 PEEP. Sxd small amts thick tan sputum. Last ABG acceptable/lungs bilat coarse crackles. ...
Radiology	522,279	[**2177-12-11**] 8:27 PM LOWER EXTREMITY FLUORO WITHOUT RADIOLOGIST LEFT Clip # [**Clip Number (Radiology) 102303**] Reason: REVISIONLEFT LEG [**Doctor Last Name **] GRAFT, SEVERE PAIN FINAL REPORT A lower extremity fluoro was performed without a radiologist present. 5.2 minutes of fluoro time was used. No films submitted. ...
Nursing	223,556	51 y/o M w/IPF on home O <sub>2</sub> who presented to the ED last night with worsening dyspnea. This has been slowly worsening for the past few weeks, but over 2 days severely worsened to the point where he was short of breath at rest. He also has had a cough productive of yellow blood-tinged sputum for 2 days (normally has a non-productive cough at baseline). According to his pulmonary rehab notes, he has been increasingly unable to exercise due to hypoxemia with exertion despite supplemental O <sub>2</sub> . He has also had anterior chest pain which he associates with coughing. The chest pain is not exertional. ...
ECG	209,051	Sinus bradycardia. The P-R interval is 0.15. There is Q-T interval prolongation and diffuse ST-T wave flattening, as well as continued T wave inversion in leads V2-V4, though improved. Otherwise, no diagnostic change. Clinical correlation is suggested. ...

*Continued on next page*

Table 3.2 – *Continued from previous page*

Category	Size <sup>1</sup>	Example of Sentences
Physician	141,624	Chief Complaint: Obtubdation Hypertensive Urgency Hyoerkalemia I saw and examined the patient, and was physically present with the ICU Resident for key portions of the services provided. I agree with his / her note above, including assessment and plan. HPI: 70 yo woman with a h/o ESRD, lives in [**Hospital1 723**], presented to HD today after missing several sessions of dialysis due to patient refusal. Pt has history of paranoia leading to HD and med refusal in past. ...
Discharge summary	55,177	Admission Date: [**2151-7-16**] Discharge Date: [**2151-8-4**] Service: AD-DENDUM: RADIOLOGIC STUDIES: Radiologic studies also included a chest CT, which confirmed cavitary lesions in the left lung apex consistent with infectious process/tuberculosis. This also moderate-sized left pleural effusion. HEAD CT: Head CT showed no intracranial hemorrhage or mass effect, but old infarction consistent with past medical history. ABDOMINAL CT: Abdominal CT showed lesions of T10 and sacrum most likely secondary to osteoporosis. These can be followed by repeat imaging as an outpatient. [**First Name8 (NamePattern2) **] [**First Name4 (NamePattern1) 1775**] [**Last Name (NamePattern1) **], M.D. [**MD Number(1) 1776**] Dictated By:[**Hospital 1807**] MEDQUIST36 D: [**2151-8-5**] 12:11 T: [**2151-8-5**] 12:21 JOB#: [**Job Number 1808**] ...
Echo	45,794	PATIENT/TEST INFORMATION: Indication: Endocarditis. Height: (in) 66 Weight (lb): 99 BSA (m2): 1.48 m2 BP (mm Hg): 130/46 HR (bpm): 79 Status: Inpatient Date/Time: [**2123-1-28**] at 10:13 Test: Portable TTE (Complete) Doppler: Full Doppler and color Doppler Contrast: None Technical Quality: Adequate INTERPRETATION: Findings: LEFT ATRIUM: Mild LA enlargement. ...

*Continued on next page*

Table 3.2 – *Continued from previous page*

Category	Size <sup>1</sup>	Example of Sentences
Respiratory	31,739	Demographics Day of intubation: 2 Day of mechanical ventilation: 2 Ideal body weight: 69.9 None Ideal tidal volume: 279.6 / 419.4 / 559.2 mL/kg Airway Airway Placement Data Known difficult intubation: No Procedure location: Reason: Tube Type ETT: Position: 23 cm at teeth Route: Oral Type: Standard Size: 7mm Tracheostomy tube: Type: Manufacturer: Size: PMV: Cuff Management: Vol/Press: Cuff pressure: cmH <sub>2</sub> O ...
Other notes	26,988	<p>[Consult Category] Respiratory failure, acute (not ARDS/[**Doctor Last Name **]) Assessment: Intubated, on CPAP [**11-23**]. Lung sounds mostly clear, required suctioning approx every 4 hours. Action: Lasix gtt continues @ 2mg/hour. Response: Patient tolerating well, approx 1.5 liters negative on [**10-18**]. Plan: Continue to diurese, attempt to wean PS and PeeP today. Provide support to patient and family. ...</p> <p>[Nutrition] Potential for nutrition risk. Patient being monitored. Current intervention if any, listed below: Comments: pt screen per icu protocol, pt currently tol pos, noted pt with some skin impairment, will sent supplements. please page if has ? ([**Numeric Identifier 1550**]) ...</p> <p>[Social Work] Pt known to this worker from prior admissions. Working with wife as she receives news of pt s critical condition and decides with team to transition pt to CMO. Wife requested that this worker meet with her and her 12 yr old son to tell him of the pt s pending death. Consulted with son s psychiatrist before meeting. Supported pt s extended family throughout the day. Pt comfortable and surrounded by family and friends ...</p>

<sup>1</sup> number of document

### 3.2.2 Sentence Boundary Detection

The text corpus is selected from the public source MIMIC-III as a case study and the data is accessed on Apr 25, 2016. The boundary detection is the comparatively fundamental task in Natural Language Processing, but significantly important regarding the text quality. In general, boundary detection task aims to detect the beginning and the ending points within given texts that a drug and a symptom have possibly participated. The challenges of boundary detection task [117, 118, 119] are arisen based on a boundary of interest and a domain of a given text. Many previous research works define a potential boundary of entity pairs candidate  $(e_i, e_j)$  within the same sentence, further, the sentence boundary detection in medical texts is also recognized as the challenge with noise-prone. One of the major issues is an ambiguous use of a *period* or a *full stop* (“.”). Typically, the *period* can be seen as a sentence boundary marker, a floating-point marker (e.g. “0.08”, “40.5 mg”), a marker for a numeric bullet of an enumerated list or a separator within an abbreviation (e.g. “y.o.”, “h.s.”), etc. The discontinuous text over a line is an irritated ill-form as well. The capital letter along with the punctuation mark such as the *colon* (“:”) for content section expression also increased the challenges. An example of noisy medical textual data is depicted in Table 3.3. The text inside the bracket pairs `[*...*]` represents the de-identification information that can be hospital name, doctor name, patient name, etc.

In this dissertation, the noise-prone is handled by developing an in-house sentence boundary detection rather than utilizing a *state-of-the-art* method to compatible with the narrative text from EMR. The heuristic patterns are predefined to carry out the unregulated linguistic form. Figure 3.2 exhibits the total number of sentences over each note section. While the number of total sentences of brief hospital course section is expressed in the highest numbers, the total number of sentences of the history of present illness section and discharge medications section are equally same. However, narrative text in Discharge Medications section mostly contains a list of prescribed drugs during hospitalization, therefore, the text in this kind of document excludes drug-event relationship (only drug is appeared in text). Conversely, brief hospital course section and history of present illness section include long text of patient treatment,

health status, or diagnosis. In order to investigate drug and event relation from EMR, sentence boundary detection method is deployed on two sections of brief hospital course section and history of present illness section. Finally, the number of extracted sentences is around 1.6 million sentences.

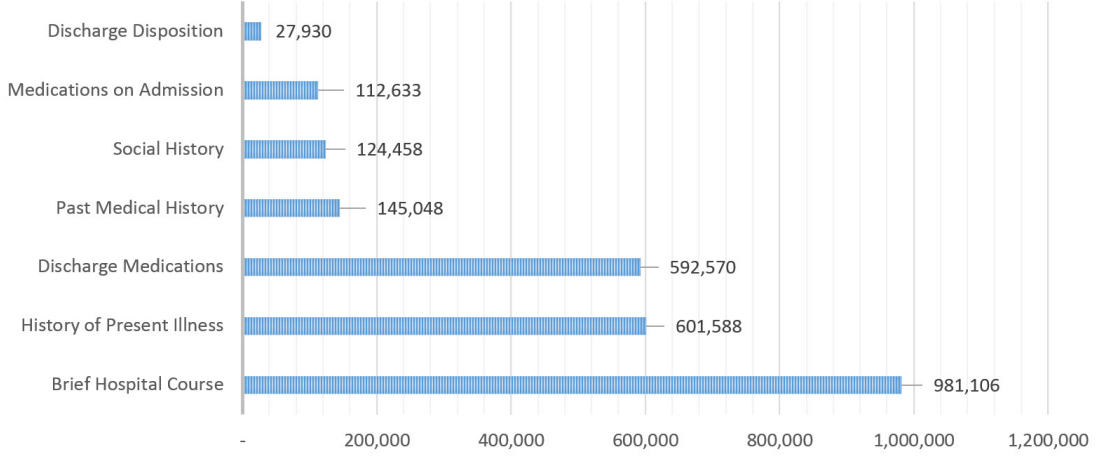


Fig. 3.2: MIMIC-III: Total number of sentences over sections in discharge summary

### 3.2.3 Medical Named Entity Recognition and Normalization

The highly accurate relation identification is strongly related to medical entities extraction. It is a common task in text mining that corresponds to named entity recognition in Information Retrieval research area. The named entity recognition is the essential subtask for information extraction to retrieve entity of interest and give its label. Given a text from Hospital Course section in MIMIC-III (see Table 3.4), suppose that we are considering on a couple of entities of a drug and a clinical event to form drug-event relation. The named entity recognition process, firstly, retrieves entities relevant the medical term of interest, then give a label to a detected medical term. Finally, we can derive information; drug entity is *codeine* and event entities are *tumor* and *persistent severe coughing*.

Similarly, the term normalization process as another subtask is also a vital process because most of machine learning and data mining method consider the distribution of entity for computation to capture the significant signal that associates to class label. Even terms refer to the same meaning but present in different written-style (synonym),

Table 3.3: (a) An example of a narrative note from a discharge summary in EMR system. (b) The noise-prone from a given text.

<p>(a)</p> <p>1: Patient was discharged on [**2001-7-24**].</p> <p>2: 20 y.o. healthy male presented to ED with worsening <b>SOB</b>. Pt</p> <p>3: states that for the past 5 days.</p> <p>4: ...</p> <p>5: Discharge Condition:</p> <p>6: ...</p> <p>7: 3. Pleuritic chest pain: Pt had peak positive Troponin of 1.45,</p> <p>8: CK 580, CKMB 10 on admission. EKG with inferolateral TWI. During</p> <p>9: his MICU course they trended down. On [**5-15**] (day prior to</p> <p>10: discharge) showed troponin of 0.08, CK of 141, CK-MB of 2. The</p> <p>11: cause was unlikely ACS and more likely myocarditis/pericarditis.</p> <p>12: Urine cocaine screen was negative. Treated pain with ibuprofen</p> <p>13: and tylenol as needed (patient uncomfortable with offers for</p> <p>14: pain control with narcotics).</p> <p>15: ...</p>	<p>(b)</p> <p>Noise-prone in medical text</p> <p>[Abbreviation]</p> <p>20 y.o. (line 2), Pt (line 2, 7)</p> <p>[List item]</p> <p>3. (line 7)</p> <p>[Floating point marker]</p> <p>1.45 (line 7), 0.08 (line 10)</p> <p>[Section boundary]</p> <p>Discharge Condition: (line 5)</p>
--	--

Table 3.4: An example of a partial narrative notes from MIMIC-III. The drug and event entities are expressed in bold.

A bronchoscopy with the intention of coring out **tumor<sub>event</sub>** was carried out by Dr.[\*\*D01-45\*\*], but all the **tumor<sub>event</sub>** was extrinsic to the airway and he was unable to relieve the obstruction. The **tumor<sub>event</sub>** now involves the trachea as well as the right main bronchus. His major complaint was of **persistent severe coughing<sub>event</sub>** and secretions. Ultimately, only **codeine<sub>drug</sub>** at 30 mg q6h controlled him and this was very affective. Inhalers provide only mild relief. He is aware of his prognosis. The patient was also seen by Dr.[\*\*D08-27\*\*] of the Oncology Service who did not feel that chemotherapy had anything of promise to offer.

it can lead to incorrect distribution. Moreover, the normalized term can support diagnosis and interpretation of professional clinicians. Table 3.5 expresses two narrative notes that refer to the same clinical events of *atrophoderma vermiculatum* (Figure 3.3<sup>2</sup>), which is primarily presenting in children with a reticular pattern of skin atrophy on the cheeks and may extend to the ears and forehead [120, 121]. The narrative notes 1 and 2 describe the patient’s condition using the terms *honeycomb atrophy* and *folliculitis ulerythematosa reticulata* respectively, however, both patients present the same disorder of *atrophoderma vermiculatum*.



Fig. 3.3: The phenotype of atrophoderma vermiculatum disorder

<sup>2</sup><http://www.dermis.net/dermisroot/en/35235/image.htm>

Table 3.5: An example of partial narrative notes from MIMIC-III. The medical terms (in bold) present the same disorder.

<p><b>Partial narrative note 1:</b></p> <p>This 8y.o. girl presented <b>honeycomb atrophy</b><sub>event</sub> on the left cheek for the past 3 years. She states that it started from the middle of cheek and extended to a larger area near ear. The patient denied the symptom presents in her family.</p>
<p><b>Partial narrative note 2:</b></p> <p>13yo. girl. On day of discharge she developed an eruption on her face, possibly <b>folliculitis ulerythematosa reticulata</b><sub>event</sub>. She states that the symptom had been present for the past 4 years. It began as a slight roughness and redness on her right cheek.</p>

The normalization intends to unify a discovered medical term into a conventional lexicon based on an identical semantic meaning or a *concept* that can be referred using concept unique identifier (CUI). There are many endeavors to deal with medical named entity recognition and normalization in medical texts such as cTAKES<sup>3</sup>, FreeLing-Med for Spanish and English, MetaMap<sup>4</sup>, tmChem<sup>5</sup>, DNorm<sup>6</sup>, GATE<sup>7</sup>, and self-developed software using CRF, parse tree, or Stanford CoreNLP tool<sup>8</sup>.

Considering on narrative notes in Table 3.3, named entity recognition can retrieve three drugs, i.e., *Ibuprofen*, *Tylenol*, *Narcotics* and six clinical events i.e. *SOB*, *pleuritic chest pain*, *ACS*, *Myocarditis*, *Pericarditis*, *Pain*. Then, the normalization task replaces a drug term or an event term with semantic concept defined by CUI. In this case, a term *Tylenol* (a trade name) is replaced with *C0000970*, which refers to a concept of *Acetaminophen* (a generic name)<sup>9</sup>, or a term *pleuritic chest pain* is replaced with

<sup>3</sup><http://ctakes.apache.org>

<sup>4</sup><https://metamap.nlm.nih.gov>

<sup>5</sup><https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/tmChem.html>

<sup>6</sup><https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/DNorm.html>

<sup>7</sup><https://gate.ac.uk>

<sup>8</sup><http://stanfordnlp.github.io/CoreNLP>

<sup>9</sup>RxNORM – <https://rxnav.nlm.nih.gov>

*C0008033*, which refers to a concept of a disorder characterized by marked discomfort sensation in the pleura<sup>10</sup> etc.

In this dissertation, the notable MetaMap tool [81] is utilized to accomplish named entity recognition. The tool recognizes a medical term from a given narrative text and results in the standard medical term corresponding UMLS. The post-processing is employed on results from *the out-of-the-box* MetaMap to overcome the ambiguously named entities. Typically, MetaMap is able to recognize multiple semantic types, a two-hierarchy of group and type of entity of interest. Therefore, in this dissertation, two semantic groups of *CHEM* and *DISO* are considered for a drug and a clinical event entities respectively. Finally, the expected result of data preparation is a normalized named entity in the form of CUI.

The summary of the statistical number of MIMIC-III corpus after data preprocessing is placed in Table 3.6. The maximum number of sentences is located in the brief hospital course section (see also Figure 3.2). On the one hand, the history of present illness section and the discharge medications sections contain the equal number of sentences; however, contexts in the discharge medications are mostly written as a list of drug prescription regardless entity description. Unfortunately, only some sections of medical notes describe the purpose of drug prescription or adverse reaction from drug usage, which can contribute to further research of drug-disease network. Therefore, this dissertation initially investigates on the brief hospital course section and the history of present illness section sections that their information is closely related to adverse drug reaction and therapeutic indication.

From the Table 3.6, the brief hospital course section contains the number of sentences more than the history of present illness section around 1.3 times and the average sentences per document are 26 and 12 respectively. However, the number of drug and event terms of the both sections exhibit as equally. It is because the history of present illness section is permeated with clinical contents that are directly related to patient’s clinical event and remedy, while, the brief hospital course section narrates patient health status before, during, and after admission including treatment courses in more details. The number of relational tuples  $(e_1, \text{pattern}, e_2)$  extraction by OpenIE are

---

<sup>10</sup>CTCAE v4.0 – <https://ctep.cancer.gov>

nearly 6.4% (77,652). In contrast, the remaining (93.6%) of extracted relational tuples contain only a drug, only a clinical event, or not related to a drug-event relation. Finally, nearly 1.7% (1,321) from 77,652 drug-event relations are corresponding known relations from SIDER and DrugBank.

### 3.3 Evaluation Metrics

#### 3.3.1 Evaluation Metric for Drug-Event Association Analysis

The performance of the proposed drug-event association method is evaluated by the lift metric which measures the degree of association by statistical analysis. In this dissertation, the lift metric is computed to assess the likelihood of a drug-event pair against the co-occurrence by chance. The value of lift over than 1 implies the stronger the association between drug and event over the chance (lift = 1).

$$lift(drug, event) = \frac{P(drug, event)}{P(drug)P(event)} \quad (3.1)$$

#### 3.3.2 Evaluation Metric for Drug-Event Identification

In order to estimate the performance of the proposed identification method, the *hold-out evaluation* is conducted through the  $k$ -fold cross-validation whereas  $k = 5$ . Accordingly, four parts of the data are used for training and the remaining one is the hold out for the validation process. Subsequently, the similar manner of data partitioning is manipulated on the next iteration with a strict rule that a validation set in the current iteration has never been used as the validation set in the previous iteration. Each iteration, eventually, would be divided into a different segment of the training and validation set. The three main measures; precision, recall and F1, are used for model evaluation (see Eq.3.2, Eq.3.3 and Eq.3.4).

- **True positive (TP)** is the number of a predicted outcome is positive and the actual is positive (correctly predict)
- **True negative (TN)** is the number of a predicted outcome is negative and the actual is negative (correctly predict)

Table 3.6: The statistical number of narrative notes from MIMIC-III after data preparation.

Description	Total (number of data)	Brief hospital course (number of data)	History of present illness (number of data)
Discharge summary	49,271	36,907	49,092
Sentences	1,580,628	980,795	599,833
Sentences with drug entity	71,201	28,186	43,015
Sentences with event entity	78,792	30,046	48,746
Sentences with drug-event pairs	218,135	124,074	94,061
Sentences/document			
+ min/max	1/251	1/248	1/118
+ avg./std.	31/23.7	26/19.1	12/8.7
Drug entity	5,825	3,141	2,684
Event entity	24,474	12,548	11,926

- **False positive (FP)** is the number of a predicted outcome is positive and the actual is negative (Type I error)
- **False negative (FN)** is the number of a predicted outcome is negative and the actual is positive (Type II error)
- **Precision** is the ability of model that correctly predicts. (see Eq. 3.2)
- **Recall** is the ability of model that is able to correctly retrieve information from the actual positive. (see Eq. 3.3)
- **F1** is the ability of model by considering the combination of precision and recall. (see Eq. 3.4)

$$precision = \frac{tp}{tp + fp} \quad (3.2)$$

$$recall = \frac{tp}{tp + fn} \quad (3.3)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (3.4)$$

## Chapter 4

# Distant Supervision-Based Pattern Bootstrapping for Relation Extraction

*In this chapter, I present the study of relation extraction, the sub-area of information extraction, to support Pharmacovigilance research community. I examine an adverse drug reaction and a therapeutic indication as semantic relations between a pair of drug and clinical event from clinical textual data in EMR as a case study. In this chapter, I make use of distant supervision on knowledge base to guideline the initial process for pattern bootstrapping. Moreover, I propose a key phrasal pattern generation by considering grammatical dependency through OpenIE, and scoring such key phrasal patterns by automatic exploring semantic relation distribution. I describe how to leverage distant supervision and a bootstrapping method to identify a semantic relation of a new pair of drug and clinical event. The content of this chapter is drawn from Taewijit, S. and Theeramunkong, T. (2016) [98].*

### 4.1 Introduction

A mediated link between two named entities indicates not only a relationship between such couple entities but also their semantic relation. For example, a sentence “*Propofol caused mild hypertension to 95*” contains a relationship between two named entities of

propofol (a drug) and hypertension (a clinical event). From the mentioned sentence, a semantic relation between propofol and hypertension relationship is adverse drug reaction (ADR). In other words, propofol possibly results in unintended hypertension. Based on such knowledge, a medical doctor should consider the adverse reaction when prescribes propofol to a patient. To achieve the relation extraction task, particularly in the medical domain, it is essential to move beyond the simplistic bag-of-words model. One of the promising techniques is the argument structure analysis that can provide information to benefit to semantic comprehension or interpretation of a sentence.

There are multidisciplinary approaches have been extensively studied to detection and identification of underlying drug-event relations. The traditional approach, statistical co-occurrence analysis is a favor for a decade [47, 46, 45] due to simplest and less effort. The method relies on simply the counting of paired drug and event entities within a specified boundary such as a window size, a sentence, an abstract, a paragraph, or a document. Unfortunately, this method loosely captures the true semantic relation such adverse drug reaction or indication.

The two complementary semantic relations; adverse drug reaction and therapeutic indication, has been recognized as significant to the comprehensive drug-disease network. Wang et al. [122] incorporate omics data, i.e., chemical structures and protein targets, and such two complementary semantic relations. In their work, adverse drug reaction and indication relations are interchangeable as a feature representation for themselves predictive model, for example, adverse drug reaction with omics data as feature representation to predict indication, and *vice versa*. Then two interdependent models are constructed to estimate the probability of indication and drug reaction relations respectively using logistic regression. Recently, the preliminary study of adverse drug reaction and therapeutic indication on social media, Segura-Bedmar et al. [123] employs statistical co-occurrence analysis of drug-event pairs by varying  $n$  window size from 10 to 50 on 400 Spanish user comments. Hence, a semantic relation is assigned based on an appearance of drug-event pairs according to its sections describing (adverse reaction or indication) that is derived from drug package leaflets.

Lately, two independent works on adverse drug reaction identification and drug indication identification are reported by Xu et al. The former work of the authors

[124] aims to derive new drug therapeutic indication for drug repurposing by exploring drug-event pairs from 20 million MEDLINE abstracts. The authors deploy pattern bootstrapping to learn a drug indication pattern from the existing data and use such patterns to detect uncover drug-event pair. Firstly, drug-event pairs are extracted from unstructured text located in Clinicaltrials.gov, the healthcare information web-based that is maintained by the National Library of Medicine (NLM). Then, the authors examine contexts between all extracted pairs to form drug indication pattern. Similarly, the latter work of the same authors [125], a dependency parse tree is deployed to retrieve adverse drug reaction specific syntactic patterns in order to discover hidden adverse reaction from the text. A large-scale of 119 million MEDLINE sentences are investigated, and the authors utilize information from SIDER knowledge base to obtain drug-event pairs corresponding adverse reaction. Next, all extracted patterns are ranked based on their associated pattern scores and co-occurrence frequencies. Finally, the manual selection process is manipulated to remove irrelevance patterns and used to retrieve unknown drug-event pairs. The example patterns from the two works of Xu et al. are such as *induced*, *associated*, *related*, *etc.* and *in*, *for-the-treatment-of*, *in-the-management-of*, *etc.* for adverse drug reaction and indication relations respectively.

According to the incredibly time-consuming, expensive, unavailable or infeasible to hand-label large amounts of training data, pattern bootstrapping method is presented as an alternative approach to learning from a few seed information and vast amounts of easily-obtained unlabeled data. In this dissertation, I study how to leverage a bootstrapping method for iterative phrasal pattern-based learning for relation extraction. Pattern bootstrapping is a general framework which incorporates unlabeled data for improving a learner. Multidisciplinary bootstrapping methods differ in how they generate a set of patterns (seed instances), represent a pattern, derive a new pattern and quantify the efficacy of a derived new pattern. Typically, pattern bootstrapping method for relation extraction initializes a learner using a few patterns. The method leverages existing patterns to generate new candidate patterns. The set of candidate patterns generated in the previous step is quantified using adjusted conditional entropy and hypothesis testing. The top rank of candidate patterns will be combined with the existing patterns and used to extract a new set of candidate patterns in the

next iteration. Finally, the process will be terminated, if a certain stop criterion is met.

Even the bootstrapping method benefits to information extraction on a large-scale unlabeled data, but its drawbacks still have room for improvement. Firstly, the initial seed method usually can be derived by manually. Although a few initial seed is proved to be successful with bootstrapping method, the bigger number of seed tends to be better, especially, when patterns of written-style are diverse such as text from EMR or social media. While biomedical literature has less distribution of writing patterns because it is carefully written and offline, EMR data has more diversity of writing pattern due to real-time records, and social media text is probably extremely various patterns. The manual seed selection is infeasible when dealing with a large amount of data, therefore, most work uses a small sample as an initial seed.

Another limitation, the initial pattern can be acquired by manually such as regular expression or rule [126], or obtained by feature representation such as TF-IDF[127]. The rule-based pattern has many advantages such as declarative, support comprehension, feasible to maintain and incorporate domain knowledge, but the main disadvantages are heuristic and laboring task [128]. Moreover, the manual pattern generation sometimes leads to ambiguity of semantic relation. The evidence is found in the work of Xu et.al.[124, 125], the term *in* is shown as the qualified pattern for both semantic relation of adverse drug reaction and indication.

## 4.2 Idea and Contributions

In this dissertation, the initial seed selection is introduced by distant supervision and pattern generation is benefited from Open Information Extraction. Distant supervision has many dominant characteristics; (i) inexpensive, no require human labor, (ii) ubiquitous, (iii) the precision of distant supervision depends on the source of knowledge base, (iv) feasible for large-scale of data. Knowledge base is ubiquitous, for instance, FreeBase and Wikipedia can benefit to person-location or person-organization relation [129], UniProtKB database can deliver the protein-locations relation [130] etc.

Regarding adverse drug reaction and therapeutic indication, SIDER and DrugBank databases are investigated for the knowledge base. An initial seed for bootstrapping

method is suggested by distant supervision. The number of seed can be derived as much as available in knowledge base, and it benefits to a size of coverage pattern. As the iterative method, the pattern construction for deriving a new set of seed is based on grammatical dependency rather than feature representation such as TF-IDF or rule base such as regular expression. The constructed pattern is called the key phrasal pattern, which is the key context around a pair of entities that can imply to the semantic relation. The key phrasal pattern is identified through open information extraction and distributional analysis. Therefore, a sentence can be simplified in the form of a relational tuple (a drug, a phrasal pattern, an event) or (an event, a phrasal pattern, a drug). Moreover, the main characteristic of a key phrasal pattern is more abundant semantic than only a verb phrase or a preposition. For instance, a key phrasal pattern *was-held-in* is more completely semantic than *held*, or a key phrasal pattern *well-controlled-with* is also more completely semantic than *with* etc. However, to reduce a specificity of the phrasal pattern, a surface lexical-based phrasal pattern is generalized using lemmatization. For example, a sentence *His outpatient ramipril was held in the setting of acute renal insufficiency*, a drug, a key phrasal pattern and an event can be derived by named entity recognition and syntactic analysis (see section 4.3) as follows.

- Drug: ramipril
- Event: acute renal insufficiency
- Surface lexicon-based phrasal pattern: was-held-in
- Syntactically lemmatized lexicon-based phrasal pattern: be-hold-in

Different from general pattern bootstrapping method, in this dissertation, the proposed bootstrapping method can identify multiple relation labels in the same process. The proposed framework of phrasal pattern bootstrapping by distant supervision has six steps as the following.

**Step1 Initial seed:** To initially generate seed by projecting drug and event pair elements from SIDER and DrugBank to a set of sentences from EMR and also sentences labeling that will be used later in Step 3.

**Step2 Construct a phrasal pattern:** A phrasal pattern is derived by considering on grammatical dependency among surrounding context, a drug and an event entities using Open IE and it returns the lemmatization of surface lexicon-based phrasal pattern as the final result.

**Step3 Rank a phrasal pattern:** A set of phrasal patterns in the previous stage are ranked by adjusting conditional entropy to investigate the distribution of uncertainty through semantic distribution. The qualified phrasal pattern is called a key phrasal pattern and be stored in a pattern pool.

**Step4 Derive new seed:** A new set of updated key phrase patterns in a pattern pool is used to extract a set of new drug-event pairs.

**Step5 Iterative pattern bootstrapping:** Repeat the process Step2 through Step5 to extract new pattern and new seed.

**Step6 Stopping criteria:** The process is terminated, if there is no more extracted new drug-event pair or no more new extracted a key phrasal pattern.

The processes in bootstrapping method can be separated into two subprocesses; relation candidate generation between a drug and an event entities and the confirmation of relation association based on the semantic of key phrasal patterns. Whereas, the final goal is to identify the plausible semantic relation of a pair of drug and event. Figure. 4.1 depicts the bootstrapping method and the whole process is exhibited in Figure.4.2.

## Contribution

I introduce the utilization of a set of outputs from OpenIE, which is domain independent information extraction, for generating a set of key phrasal patterns. Moreover, the generated key phrasal patterns are represented by syntactically lemmatized lexicon-based phrasal pattern that overcomes the semantic drift issue in the learning process of a bootstrapping method. Generally speaking, the meaning between a pair of drug and event is retrained through multiple iterations. In addition, I propose a method to automatically quantify an efficacy of a key phrasal pattern corresponding binary label of ADR and IND rather than human labor [124, 125]. I contribute how to leverage a bootstrapping method to identify the semantic relation of two entities of drug and

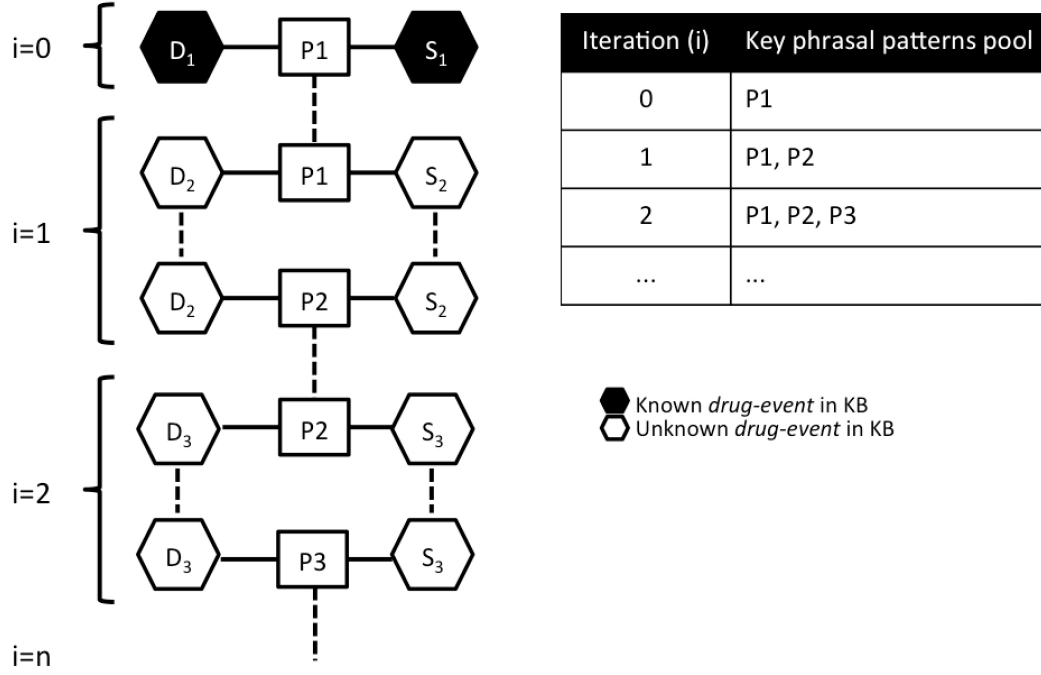


Fig. 4.1: The bootstrapping method for drug-event relation extraction.

event which are appeared in the same sentence as a case study.

### 4.3 Proposed Method

The biomedical relation extraction in text mining in this chapter involves three main tasks; (i) initial seed generation, (ii) phrasal pattern generation, (iii) bootstrapping and (iv) semantic relation inference. For the first task, it is related to the generating of potential entity pair candidates that have a tendency to form a relationship that is supervised by distant supervision. The second task is involved to automatic pattern generation by considering grammatical dependency among drug and event entities. The third task is an iterative process to derive new seed and pattern including confirmation of the association between such potential drug-event candidates. The last one aims to infer the semantic relation or assign label base on the statistical association. Algorithm 1 describes pseudo-code of the proposed bootstrapping method and Figure 4.2. illustrates the overall of the proposed method. The data preprocessing is deployed on narrative notes in EMR using the method that is described in Section 3.2.2 and 3.2.3.

The two knowledge bases, SIDER and DrugBank, are comparative enrichment as seed generation instead of manual selection by a domain expert. The parsing is deployed in order to retrieve a set of phrasal patterns for a given set of seed in the previous stage. The distributional semantic relations across derived phrasal patterns are examined and qualified by adjusted conditional entropy. Lastly, the inference process is employed to derive novel semantic of a pair of drug and event.

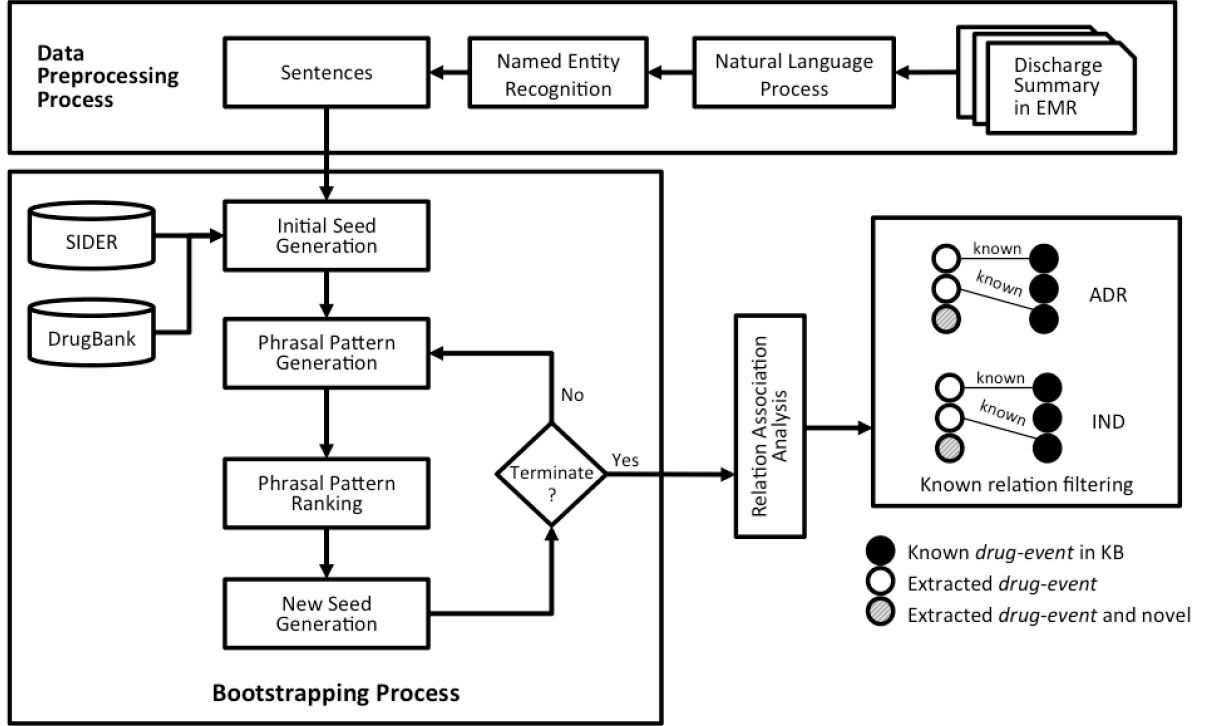


Fig. 4.2: The proposed method for phrasal pattern bootstrapping for semantic relations identification.

#### 4.3.1 Distantly Supervised Initial Seed

A pair of drug and an event can be initially guided by the fact from knowledge bases. The method is promising and feasible for large-scale of unlabeled data. This method can be thought as manual pattern generation. Firstly, domain expert selects a set of sentences that contain both of a drug and an event of interest, then considers on the context around drug-event entities and generates patterns to represent the semantic relation of drug-event pairs. Even the starting process needs data preprocessing that

---

**Algorithm 1:** Pseudo-code for the proposed bootstrapping method

---

**Input:** $\mathcal{K}$  = knowledge bases $\mathcal{X}$  = a set of sentences $S_{all} = \emptyset$ ; a set of seeds $P_{best} = \emptyset$ ; a set of the best patterns $P_{candidate} = \emptyset$ ; a set of candidate patterns $T$  = the maximum number of iteration**Output:**  $P_{best}$ **1**  $t \leftarrow 0$ **2**  $S_{all} \leftarrow InitialSeed(\mathcal{K}, \mathcal{X})$ **3 repeat****4**      $P_{candidate} \leftarrow ExtractPattern(S_{all})$ **5**      $P_{best} \leftarrow Score(P_{candidate})$ **6**      $S_{all} \leftarrow QueryNewSeed(P_{best})$ **7**      $t \leftarrow t + 1$ **8 until** *convergence or*  $t = T$ 

---

is involved in natural language process, the projecting process of fact from knowledge bases to each sentence is simple.

In this process, the fact from a database will be projected to a given sentence and also its label that corresponds to a semantic relation of the fact, and automatic pattern generation will be described in the next section (see Section 4.3.2). For example, given a sentence *Metoprolol was discontinued due to hypotension*, the algorithm looks up a pair of drug and event of *metoprolol-hypotension* in SIDER database. If the fact expresses the true relation of such drug-event pair, the ADR label is assigned into all sentences that contain *metoprolol* and *hypotension*. Similarly, given another sentence *Labetalol was ordered for his hypertension*, the algorithm looks up a pair of drug and event of *labetalol-hypertension* in DrugBank database. If the relation of such drug-event pair is found in DrugBank database, the IND label is assigned into all sentences that contain *labetalol* and *hypertension* as well.

### 4.3.2 Automatic Phrasal Pattern Generation

Surrounding context among a pair of drug and event entities can imply the semantic relation between such drug-event pair. The representation of such surrounding context or pattern can be a representative feature, regular expression or rule base. This dissertation introduces automatic pattern generation by considering syntax tree or parse tree. The advantage of the parse tree for information extraction is well-known as to preserve the structure of natural language and provide semantic relation for human interpretation and comprehension. However, the main weakness of traditional information extraction is domain dependency. Recently, Open Information Extraction (OpenIE), which is a generalization of typical information extraction, is introduced to overcome such limitation. OpenIE provides the potential effort to deal with large-scale corpora without manual tagging of relations [86], while the traditional IE fully requires precisely target relation beforehand. Early of OpenIE [131, 132] aims to extract an unknown relation in advance on highly scalable Web corpus. The evident achievements on web mining lead to an extensive paradigm shift in medical text mining. Recently, the Stanford CoreNLP developed OpenIE tool [92] to reduce a large pattern set for

Table 4.1: The statistical number of extracted relations derived by OpenIE from narrative notes in MIMIC-III.

	MIMIC-III			Knowledge base (KB)		
	Total	BHC <sup>1</sup>	HPI <sup>2</sup>	Total	BHC <sup>1</sup>	HPI <sup>2</sup>
All open relations	1,210,501	675,664	534,837	-	-	-
drug terms	1,142	639	977	192	168	78
event terms	3,080	2,143	2,111	190	171	78
drug-event pairs (ADR)	77,652	43,088	34,564	589	480	109
drug-event pairs (IND)				732	553	179

<sup>1</sup> The brief hospital course

<sup>2</sup> The history of present illness

canonical sentences and excerpt self-contained clauses from longer sentences as well.

In this research, given a set of medical textual sentences, the Stanford OpenIE is carried out to examine the powerful on clinical text mining (Figure 5.2). The upper block depicts the dependency parsing of two sentences (S1 and S2) and their outputs from OpenIE. The lower table exhibits their final representations in the form of a relational table. Generally speaking, this syntactic-based analyzer extracts a list of domain-independent relational form ( $arg_1$ , pattern,  $arg_2$ ) from the sentences, and a subset of such massive numbers of domain independent relation corresponds to *drug*, *key phrasal pattern*, *event* tuples (d, p, e), where drugs and events are matched with their corresponding CUIs.

As the results, 1.2 million of the massive numbers of the domain-independent relational form ( $arg_1$ , pattern,  $arg_2$ ) are reported. As mentioned above, the only subset of relation outputs can represent a drug-event relationship. Hence, the post-processing is manipulated on outputs from OpenIE. The irrelevance relation outputs which  $arg_1$  and  $arg_2$  are not represented by a named entity of a drug and an event (or an event and a drug) are filtered out. Subsequently, 77,652 of drug-event pairs candidate are generated. Table 4.1 describes the statistical number of relations derived by OpenIE.

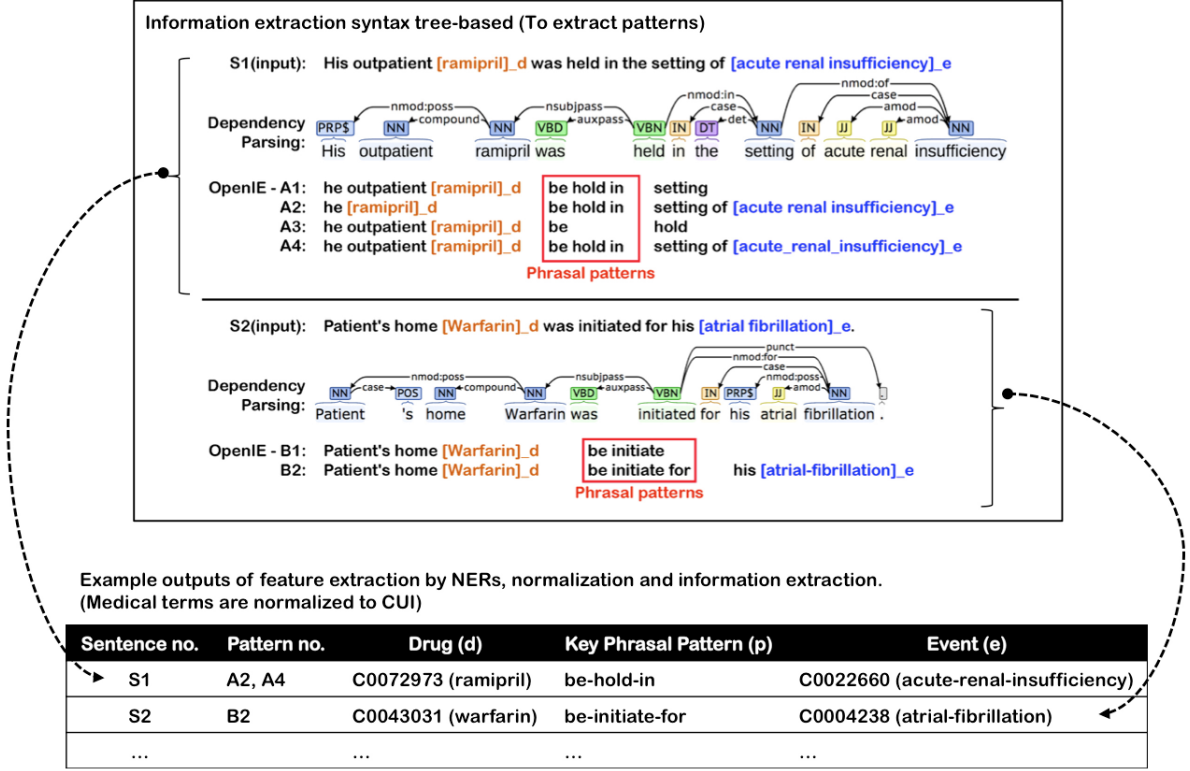


Fig. 4.3: Open Information Extraction for given two medical sentences from EMR.

### 4.3.3 Phrasal Pattern Scoring

Regarding phrasal pattern generation derived by considering on the context around a pair of drug and entity in a sentence. A qualified phrasal pattern should express the high degree of discriminative ability between such binary semantic relations. In the other words, a qualified phrasal pattern should specific to its semantic relation (i.e. ADR, IND). Therefore, the distribution of a candidate phrasal pattern across each semantic relation is observed. The conditional entropy as shown in Eq.4.1 is examined to quantify the degree of uncertainty for each phrasal pattern. After that, the phrasal pattern strength is obtained by conditional entropy adjustment as shown in Eq.4.2. The higher score implies the stronger phrasal pattern strength. Finally, unreliable phrasal patterns are excluded by the hypothesis testing of association between the semantic relation given a candidate phrasal pattern. The statistical Fisher's exact test at 0.05 significant level is considered. The qualified phrasal pattern in this phase is so-called *a key phrasal pattern*.

Given a candidate phrasal pattern  $x_i$  and semantic relation  $y_j \in Y$  whereas  $Y =$

$\{ADR, IND\}$ , the conditional entropy and pattern strength are defined as follows:

$$H(Y|X = x_i) = - \sum_{j=1}^C P(y_j|x_i) \log_2 P(y_j|x_i) \quad (4.1)$$

$$P_{strength}(X = x_i) = (1 - H(Y|X = x_i))(P(y_j|x_i) - (1 - P(y_j|x_i))) \quad (4.2)$$

#### 4.3.4 Iterative Seed Generation

A new set of the key phrasal patterns from the previous step is used to retrieve other a new set of drug-event pairs. The lemmatized form of a key phrasal pattern is used to retrieve a new seed. A new set of seed that can be mapped by a key phrasal pattern will be repeated using Step 2 to automatically construct a phrasal pattern. Then a set of phrasal patterns that belong to such discovered drug-event pairs will be extracted. A new set of phrasal patterns will be pooled to  $P_{candidate}$  and iterate to score and choose the significant one. The process will be terminated if no more extracted candidate pattern (converge) or the iteration is more than the specified maximum iteration.

In summary, 353 qualified phrasal patterns (216 for *ADR* and 137 for *IND*) are derived. Opposition from the previous study by Xu et al. [124, 125], the proposed method is automatic semantic relation identification, non-redundant, and feasible for a large amount phrasal patterns extraction. The ranking of key phrasal patterns is exhibited in Figure 4.4. A set of key phrasal patterns is ranked by the pattern strength (gray area), the higher score, the stronger pattern strength. The frequency of drug-event corresponding ADR and IND are exhibited with red and green lines accordingly. Additionally, Table 4.2 illustrated the top 5 key phrasal patterns and sample sentences of drug-event corresponding ADR and IND semantic relations.

#### 4.3.5 Semantic Relation Inference

This phrase is to evaluate the performance of the proposed method by association analysis. A set of key phrasal patterns corresponding ADR and IND is used to retrieve drug-event pair from a set of MIMIC-III corpus in order to infer semantic relation or assign relation label. A subset of drug-event pairs is randomly selected to investigate

Table 4.2: Top 5 of the extracted key phrasal pattern and the example sentences from MIMIC-III.

Top 5	Key phrasal patterns	Example sentences
<i>drug-event</i> (ADR)		
1	be hold in	<u>His outpatient ramipril was held in the setting of acute renal insufficiency</u>
2	contribute to	<u>Morphine contributed to urinary retention as seen in high PVR so foley placed</u>
3	be think	<u>His rash was thought secondary to the nafcillin</u>
4	improve with	<u>Patient's dysphagia improved with iv pantoprazole</u>
5	cause	<u>Propofol caused mild hypotension to 95</u>
<i>drug-event</i> (indication)		
1	continue	<u>Depression - continue outpatient fluoxetine</u>
2	be start for	<u>Phenylephrine drip was started for hypotension</u>
3	be on	<u>RHEUMATOID ARTHRITIS: The patient is on Methotrexate at home</u>
4	be control with	<u>The patient's blood pressure was controlled with labetalol</u>
5	be add for	<u>Norepinephrine was later added for persistent hypotension</u>

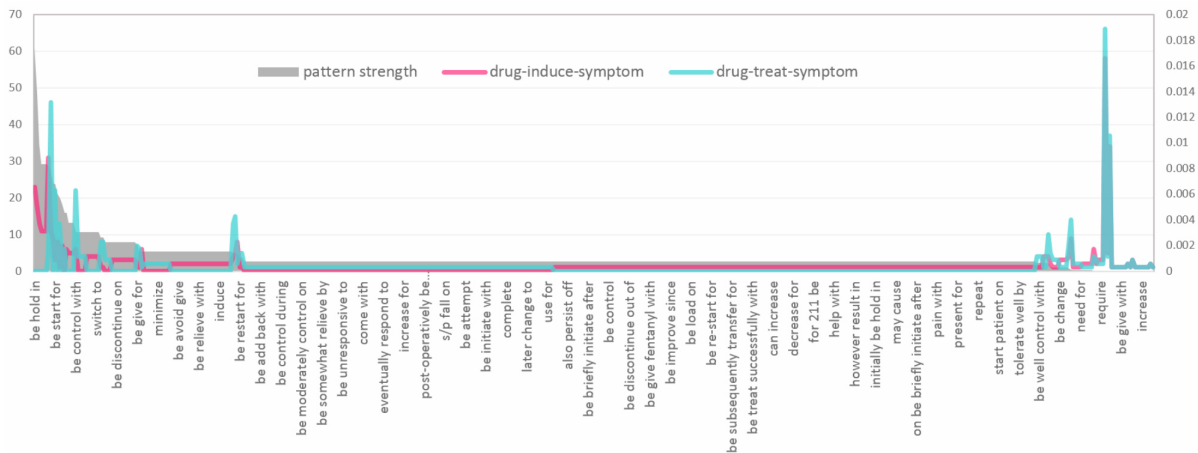


Fig. 4.4: Plot of discovered key phrasal patterns across frequency and pattern strength.

statistical association. Then the lift metric (see Section 3.3.1) is computed to evaluate the likelihood of a pair of drug and event against the co-occurrence by chance. The lift value over than 1 implies the stronger association between a drug and an event over the chance ( $\text{lift} = 1$ ). The experimental results will be discussed later in Section 4.4.3.

## 4.4 Evaluation

The proposed method is evaluated and reported in three parts; (i) analysis of extracted key phrasal patterns, (ii) evaluation on the effectiveness of the key phrasal pattern and (iii) evaluation on the discovered drug-event pair by domain experts.

### 4.4.1 Data

The proposed pattern bootstrapping method is involved to two datasets. As the first dataset, the fact from knowledge bases from SIDER and DrugBank is acquired for automatic data curation by distant supervision. While a set of pairs of drug-event in SIDER database is used for ADR relation, a set of drug-event pairs in DrugBank database is used for IND relation. Fortunately, both databases are publicly available at [sideeffects.embl.de](http://sideeffects.embl.de) and [www.drugbank.ca](http://www.drugbank.ca) respectively. Another dataset, the medical text corpus, nearly 1.6 sentences 50,998 discharge summary derived by the preprocessing process (see Section 3.2) is used in this experiment.

#### 4.4.2 Analysis of Extracted Key Phrasal Patterns

Using the iterative approach with distant supervision and grammatical dependency analysis, 353 key phrasal patterns are generated. There are 216 and 137 key phrasal patterns corresponding ADR and IND relations accordingly. Such discovered key phrasal patterns are compared with the one from the two studies of Xu et al. [124, 125].

The key phrasal patterns from the proposed method in this dissertation are different from Xu et al. due to the different writing style of two sets of the medical corpus. This dissertation uses EMR corpus, but in the work of Xu et al. uses biomedical literature as a corpus. Moreover, the dissimilar extracting pattern methods also result unlike pattern outputs. From Table 4.4, the key phrasal patterns from the proposed method benefit to ADR and IND relation extraction than the study of Xu et al. via MEDLINE corpus in multiple aspects. The comparison of the key phrasal patterns is provided as the followings and its summary is provided in Table 4.3.

**Phrasal pattern generation method:** the proposed method uses automatic dependency parsing, while, the first study of Xu et al. [124] uses manual pattern generation. The phrasal pattern construction is suggested by a domain expert, however, dealing with a large scale of medical text is infeasible for manual process due to the limitation of computational effort in human at a time. Another study of Xu et al. [125], they employ the parse tree as well, but the authors do not provide the details about how to derive their patterns.

**Length vs. comprehension of key phrasal pattern:** the proposed method provides longer length of a key phrasal pattern, but the studies of Xu et al. result the shorter one. For instance, the proposed method extracts the key phrasal pattern *be in* but the output from Xu et al. is *in*. Another example, the key phrasal pattern *be control with* is reported by the proposed method while the pattern *with* is generated by Xu et al. The longer length of the key phrasal pattern contains verb and follow by preposition, the shorter one as shown in the studies of Xu et al. contains only verb or preposition. Furthermore, the polarity of a key phrasal pattern can be found in the proposed method such as *well control with*. This pattern contains the word *well* that refers to the positive polar that can support the confidence of the pattern spe-

cific semantic relation level. In this case, the phrasal phrase *<event> well control with <drug>* expresses ADR semantic relation. Mostly, the extracted key phrasal patterns of the proposed method are fully contained *verb* and followed by *preposition*, which benefits to comprehensive meaning corresponding ADR and IND.

**Redundant phrasal pattern:** the proposed method reduces the variety of the inflected forms of a word by considering on the syntactic lemmatized lexicon. Any word in a phrasal pattern will be transformed to the base form, for example, the word *cause, causes, caused* are changed to *cause*. This method can benefit to generalization and supports the distributional semantic exploration. For the studies of Xu et al., the authors construct a pattern based on surface lexicon, therefore the redundant patterns such as *induced, induces* are treated as the different patterns in their work.

**Discriminative ability:** the first study of Xu et al. considers the binary problem of IND as IND and non-IND and the similar manner also employs to the binary relation problem of ADR in their second work. During the pattern extraction phase in both studies of Xu et al., the authors do not consider IND and ADR as the complementary problem. Generally speaking, the authors separately extract patterns as the individual task. Therefore, the same pattern such as *after* can be appeared as the key phrasal pattern to derive both ADR and IND relation. This is the extremely equivocal pattern and hazardous case because we cannot conclude the semantic relation of a derived drug-event pair by using such kind of pattern. The uncertain pattern is well-known as the cause of false positive. Different from the proposed method, there is no overlapped pattern for both semantic relations because the proposed method explores the distribution of ADR and IND during the phrasal pattern generation. One phrasal pattern can be the key pattern to represent only one semantic relation, and the degree of confidence is expressed by adjusted conditional entropy.

#### 4.4.3 Evaluation on the Key Phrasal Pattern VS. Semantic Relation Specificity

The semantic relation identification of an arbitrary pair of drug and event can be derived by the key phrasal patterns inference. All 353 extracted key phrasal pat-

Table 4.3: The summary table of the key phrasal pattern comparison.

Key Phrasal Pattern Comparison	The proposed method	The studies of Xu et al.
Phrasal pattern generation method	Parse tree, Distributional relation analysis	Manual, Parse tree
Length of patterns	Longer	Shorter
Complete phrase	More complete	Less complete
Polarity expression in patterns	Yes (some patterns)	No
Redundant pattern	Non-redundant	Redundant
Discriminative ability	High	Low

terns are employed. The proposed method successfully identifies approximately to five times increasing (6,347 drug-event pairs) from the known relation. Four sets of drug-event pairs corresponding ADR and IND are randomly selected; *Metoprolol-event* and *drug-Hypotension* for adverse drug reaction semantic relation, and *Amiodarone-event* and *drug-Pneumonia* for *indication* semantic relation. Table 4.5 exhibits the discovered drug-event relation of the four sets and their likelihood ratio. The KB column is marked *yes*, if drug-event pair exists in the knowledge base (SIDER or DrugBank). In the contrast, the KB column is marked *new* for the discovered drug-event pair by the proposed method.

From FDA prescribing information, *Metoprolol* drug is indicated to treat chest pain, hypertension, and prevents the heart attack. A key phrasal pattern relevant to ADR semantic relation is employed to retrieve *Metoprolol-event* pair. The results show that the frequent and common ADR caused by *Metoprolol* such as *AV block*, *Heart block*, *Hypotension* has the higher lift value, while the no frequent information or rare such as *rash* and *tachycardia* provided the small number over chance. The novel *Metoprolol-Pulmonary* as a drug-event pair that is not existing in the knowledge bases, are derived from the proposed method. The RxList<sup>1</sup>, which is the premier Internet Drug Index resource, is used to verify the identified semantic relation. As a result, it is found that *dyspnea of pulmonary origin* is reported as ADR of *Metoprolol* drug. Another finding, the novel *Metoprolol-Kidney failure* pair is reported by Mayo Clinic<sup>2</sup> as the possible ADR for long-term treatment. Identically, *drug-Hypotension* as a drug-event pair is

<sup>1</sup><http://www.rxlist.com/>

<sup>2</sup><http://www.mayoclinic.org/>

Table 4.4: The comparison of the key phrasal patterns between the proposed method and the two studies of Xu et al.

	ADR		IND	
	drug-event	event-drug	drug-event	event-drug
The proposed method	<i>be hold in</i>	<i>be in</i>	<i>continue</i>	<i>be continue on</i>
	<i>be hold</i>	<i>improve with</i>	<i>be start for</i>	<i>be control with</i>
	<i>contribute to</i>	<i>be think</i>	<i>be add for</i>	<i>be on</i>
	<i>be hold for</i>	<i>be attribute to</i>	<i>be initiate for</i>	<i>well control with</i>
	<i>cause</i>	<i>think</i>	<i>be give for</i>	<i>continue on</i>
	<i>be discontinue</i>	<i>feel</i>	<i>be continue for</i>	<i>be control on</i>
	<i>be hold give</i>	<i>hold for</i>	<i>be restart for</i>	<i>be treat with</i>
	...	...	...	...
	<i>_induced</i>	<i>induced_by</i>	<i>in</i>	<i>with</i>
	<i>induced</i>	<i>after</i>	<i>for the treatment of</i>	<i>were treated with</i>
Xu et al. [124, 125]	<i>_associated</i>	<i>caused by</i>	<i>treatment of</i>	<i>to</i>
	<i>_related</i>	<i>following</i>	<i>in the management of</i>	<i>after</i>
	<i>induces</i>	<i>produced by</i>	<i>_resistant</i>	<i>during</i>
	<i>caused</i>	<i>after treatment</i>	<i>in a patient with</i>	<i>in</i>
	<i>developed</i>	<i>in patients treated</i>	<i>to treat</i>	<i>associate with</i>
	...	...	...	...

examined by the proposed method. From knowledge in SIDER database, hypotension is a common ADR of *Losartan* and *Metoprolol* drugs, and as a result shown in Table 4.5, the lift value of such drug-event pair is placed in the high order. In the contrast, the *Ciprofloxacin-Hypotensions* pair has the lower lift value due to the rare and uncommon ADR.

For therapeutic indication semantic relation, the lift value is totally different from adverse drug reaction relation. The lift value of IND relation exhibits the bigger number due to the drug prescribing regularly with a known indication for hospitalization. Considering on *Amiodarone-event* as a drug-event pair, all clinical events relevant *Amiodarone* drug as shown in the lower left of Table 4.5 are related to heart rhythm disorders, and such treatment indication has existed in the DrugBank knowledge base; therefore, the lift value of all *Amiodarone-event* pairs exhibit high score. Another one, *Pneumonia* clinical event, this event is an infection of the lungs, and it can be caused by bacteria, viruses or fungi. All of the drugs relevant to *Pneumonia* in the pattern *event-Pneumonia* as shown in the lower right of Table 4.5 are indicated to treat bacterial infections. However, only *Azithromycin* and *Cefepime* drugs are reported in DrugBank database. The proposed method can derive the alternative drug therapy, i.e., *Levofloxacin* and *Ceftriaxone* drugs for *Pneumonia* with the lift score of 24.95 and 14.40 respectively.

#### 4.4.4 Evaluation on the Discovered Drug-Event Pair by Domain Experts

In order to evaluate the performance of the proposed method in practical used by a professional clinician, two domains are invited to evaluate and provide comments. The first domain expert is a lecturer and a researcher in pharmaceutical domain, and another one is a medical doctor and a researcher in hospital university. Sentences that contain drug-event pairs of drugs; *Metoprolol*, *Imdur*, *Levofloxacin*, *Amiodarone*, *Methotrexate*, are randomly selected. The domain experts evaluate and provide comments on each sentence as shown in Table 4.6.

Moreover, some sentences contain multiple drugs combination and the key phrasal

Table 4.5: The evaluation on key phrasal patterns: discovered drug-event pairs vs. knowledge base.

Relation	Metoprolol-event	lift	KB (SIDER)	drug-Hypotension	lift	KB (SIDER)
ADR	<i>AV block</i>	11.91	yes	<i>Losartan</i>	5.67	yes
	<i>Heart block</i>	11.91	yes	<i>Metoprolol</i>	5.24	yes
	<i>Pulmonary disease</i>	5.96	new	<i>Propofol</i>	4.99	yes
	<i>Hypotension</i>	5.24	yes	<i>Carvedilol</i>	4.86	yes
	<i>Asystolic</i>	3.40	yes	<i>Imdur</i>	4.86	new
	<i>Sepsis</i>	2.70	new	<i>Furosemide</i>	3.60	yes
	<i>Kidney failure</i>	2.38	new	<i>Nadolol</i>	3.24	yes
	<i>Tachycardia</i>	2.09	yes	<i>Atenolol</i>	3.13	yes
	<i>Rash</i>	1.54	yes	<i>Ciprofloxacin</i>	2.05	yes
	Amiodarone-event	lift	KB (DrugBank)	drug-Pneumonia	lift	KB (DrugBank)
IND	<i>Ventricular arrhythmia</i>	28.40	yes	<i>Levofloxacin</i>	24.95	new
	<i>Rhythm</i>	19.88	yes	<i>Azithromycin</i>	17.43	yes
	<i>Ventricular tachycardia</i>	17.04	yes	<i>Ceftriaxone</i>	14.40	new
	<i>Atrial fibrillation</i>	11.09	yes	<i>Cefepime</i>	11.20	yes

pattern might not be appropriate for making a decision on the relation labels. For example, a sentence “His pneumonia was treated with Levofloxacin and Metronidazole and subsequently Vancomycin as well (sputum grew MSSA).”. The proposed method can identify *Levofloxacin-Pneumonia* relationship as drug indication through the key phrasal pattern *be-treat-with*. The domain experts provide the opinion that pneumonia is likely treated by the combination of three medicines; *Levofloxacin*, *Metronidazole* and *Vancomycin*. From the error analysis, it is found as the limitation of OpenIE that could not retrieve the indirect patterns of *Metronidazole-Pneumonia* and *Vancomycin-Pneumonia*. Therefore, it has a room for improvement on open information extraction for the future work. However, the drugs combination and adverse reaction or the drugs combination and therapeutic indication relationship can be treated as the other problem.

Table 4.6: The evaluation by domain experts on the randomly selected sentences.

Sentence	Drug	Key Phrasal Pattern	Event	Relation Label	Comments by domain experts
Metoprolol was discontinued because of AV block on amiodarone IV initially.	<i>Metoprolol</i>	<i>be-discontinue-because-of</i>	<i>AV block</i>	ADR	Possibly happened.
Given her severe underlying pulmonary disease, the patient's metoprolol was changed to diltiazem.	<i>Metoprolol</i>	<i>be-change-give</i>	<i>Pulmonary disease</i>	ADR	Metoprolol is might be the cause of pulmonary problem especially underlying pulmonary disease, common symptom: shortness breath.
Losartan should be restarted once acute renal failure is resolved, topol was switched to metoprolol for better ability to titrate.	<i>Metoprolol</i>	<i>be-switch-to</i>	<i>Kidney failure</i>	ADR	Possibly happened.
However, there was spontaneous resolution of the rash, HTN: Metoprolol and Lisinopril continued.	<i>Metoprolol</i>	<i>spontaneous-resolution-of</i>	<i>Rash</i>	ADR	Rash is uncommon side effect but it can be occurred.
Metoprolol was discontinued because of some intermittent hypotension.	<i>Metoprolol</i>	<i>be-discontinued-because-of</i>	<i>Hypotension</i>	ADR	Possibly happened. Hypotension is common adverse drug reaction.
Imdur was stopped due to relative hypotension (SBP 90s-110s).	<i>Imdur</i>	<i>be-stop</i>	<i>Hypotension</i>	ADR	Possibly happened. Hypotension is common adverse drug reaction.

*Continued on next page*

Table 4.6 – *Continued from previous page*

Sentence	Drug	Key Phrasal Pattern	Event	Relation Label	Comments by domain experts
She was empirically treated for pneumonia with levofloxacin; although there is no compelling evidence of pneumonia or infiltrate on CXR.	<i>Levofloxacin</i>	<i>be-with</i>	<i>Pneumonia</i>	IND	Yes. Levofloxacin is used to treat pneumonia.
Amiodarone therapy was also initiated for episodes of paroxysmal atrial fibrillation.	<i>Amiodarone</i>	<i>be-initial</i>	<i>Atrial fibrillation</i>	IND	Yes. Amiodarone is used to treat cardiac problem.
Amiodarone was started due to ventricular arrhythmia in the operating room and was stopped post operative day one due to no further rhythm issues.	<i>Amiodarone</i>	<i>be-start</i>	<i>Ventricular Arrhythmia</i>	IND	Yes. Amiodarone is used to treat ventricular arrhythmia.
RHEUMATOID ARTHRITIS: The patient is on Methotrexate at home.	<i>Methotrexate</i>	<i>be-on</i>	<i>Rheumatoid Arthritis</i>	IND	Yes. Methotrexate is used to treat rheumatoid arthritis.

## 4.5 Summary

Distant supervision and iterative pattern bootstrapping framework are proposed to identify the semantic relations of ADR and IND from the large-scale of narrative text from EMR. After data preprocessing, nearly 1.6 million sentences are extracted and used in the further processes. In this chapter, the distant supervision is used to curate drug-event relation as a set of initial seed instead of a domain expert. A collection of phrasal patterns candidate is constructed using dependency parsing deployed by Stanford OpenIE. The iterative bootstrapping approach is processed to make qualification on such set of phrasal patterns candidate by exploring phrasal patterns specific semantic relations. The adjusted conditional entropy with  $0.05$  significant level is presented to capture the pattern strength and automatically qualify the candidate phrasal pattern instead of manually selection. This process results in a set of key phrasal patterns. Furthermore, the key phrasal pattern inference is employed in order to identify the hidden semantic relation of a new drug-event pair. To this end, the lift metric is computed to measure the likelihood of semantic association of a pair of drug and event. To derive novel drug-event pairs and their semantic relation, known drug-event pairs that correspond to SIDER and DrugBank databases are excluded.

The proposed method has some limitations that need to be improved such as low recall rate due to small numbers of the key phrasal pattern and the precision of a drug and an event named entity recognition task. In addition, dependency parsing from Stanford OpenIE can retrieve diversity of relational tuples, however, the method fails to discover partial or incomplete sentence especially the absent of *verb* that can be found in some parts of the narrative text in EMR.

From the experimental results, the proposed method is not only effective and scalable for semantic relation identification, less expensive for data labeling by a domain expert, but also promising framework to discover a novel harmful and beneficial drug therapeutic indication. This preliminary investigation of the utilization from EMR indicates that the contribution of this work can support the further research of drug safety surveillance and drug repurposing as a screening method by the systematic way. As the important role of distant supervision for large-scale unlabeled data, the next

chapter will introduce the improvement of adverse drug reaction and therapeutic indication identification task by using distant supervision approach for data labeling in supervised and semi-supervised learning process. The key phrasal patterns in this chapter are also utilized as the feature representation of a drug-event pair in the next chapter as well.

## Chapter 5

# Distant Supervision-Based Transductive Inference for Relation Extraction

*This chapter presents an extensive work of the previous chapter to improve the performance of ADR identification. Different from the previous chapter, here, I leverage distant supervision for ground truth construction. I propose an alternative parameters estimation for a generative model to overcome a limitation of the traditional assumption of word independence. I also perform the assessment through multiple parameters such as feature representation, weighting models, initial weightings of relations for unlabeled data incorporation and the comparison between a proposed method and advanced methods. The content of this chapter is relied on Taewijit, S., Theeramunkong, T, and Ikeda, M. (2017) [133].*

### 5.1 Introduction

The make use of distant supervision for pattern extraction in bootstrapping process is shown to be successful in the previous chapter. In this chapter, distant supervision will be used for training data construction instead of manual handcrafted labeling. There is a few work for ADR identification using distant supervision as training data construction. Table 5.1 describes the summary of previous studies on ADR identification.

Earlier research, the statistical co-occurrence method is broadly employed to quantify the relationship strength between a drug-event pair. While the method is simple, its result might present no explicit clinical relevance of a derived drug-event pair [12] due to disregard relational context that might express an exact impressive in a clinical event such as a drug treats a symptom or a drug causes a symptom. To fill in this research gap, many researchers consider surrounding contexts around drug and event entities within clinical texts and represent such data by either using pattern-based method [112, 113, 115, 98, 125, 134] or features-based method [99, 135, 34]. Consequently, a potential ADR is identified by either training supervised learning or semi-supervised learning [13] model. However, there are two main difficulties when dealing with unstructured texts using such learning models. A rare availability of labeled instances derived by human annotation to form a gold-standard example is the former problem, and intractable processing of unstructured clinical texts is the latter one. Toward the insufficiency of labeled instances, several studies alleviate this problem by using a sort of heuristics or rules (distant supervision [14, 15]); i.e., mapping a sentence that contains entity pair  $(e_1, e_2)$  from knowledge base and tagging relation label  $(y)$  to such mentioned sentence to form a training set. For the second problem, a word-based approach [20, 136], the most commonly used method for text representation, is introduced; however, the method ignores either grammatical or semantic dependency among words. Therefore, pattern-based methods [112, 113, 125] are promoted to either extensive or substitute for word-based text representation. Recently, distant supervision paradigm is introduced to overcome hand-labeled data process to obtain a label of an instance from knowledge base [14, 15]. For example, knowledge bases consist of the following drug-event relations; (“*ramipril-allergy*”, “ADR”) and (“*aspirin-fever*”, “IND”), so-called *entity-level* relation. By distant supervision, we can derive automatic labeled data of an associated sentence with such drug-event, e.g., “His *ramipril* were discontinued due to *allergy* and added to list in our medical records”, “ADR”, and known as *instance-level* relation. Therefore, multiple-instant learning (MIL) paradigm [137] is introduced into the classifier builder process to handle such two-levels relations.

This chapter introduces ADR identification framework by aiming to classify an *entity-level* relation of a drug-event pair. The proposed work differs from prior related

works in the following aspects: (i) I propose key phrasal pattern-based bootstrapping method for characterizing ADR and IND; (ii) I introduce alternative parameters learning of a generative model (iii) I perform enhancement of the proposed method by incorporate transductive learning method.

Table 5.1: A list of previous studies on ADR identification from unstructured text

Data Source	Literature	Year	Size	Label number	Labeling Method	NER	Method
<b>Supervised Learning</b>							
EMR	Aramaki et al. [112]	2010	3,012 notes	A,O (2)	H	CRF	Pattern-based
	Sohn et al. [113]	2011	237 notes	A,O (2)	H	cTAKES	Pattern-based, DT C4.5
	Henriksson et al. [114]	2015	400 notes	A,I,O (3)	H	CRF	Word embedding, RF
	Casilas et al. [115]	2016	n/a	A,O (2)	H	FreeLing-Med	Pattern-based, SVM, RF
	Peng et al. [99]	2016	18,410 abstracts	A,O (2)	H, DS	Dictionary, tmChem, DNorm	Features-based, SVM
Social Media	Segura-Bedmar et al. [97]	2015	84,000 messages	A,I (2)	DS	GATE	Shallow linguistic kernel, Distant supervision
	Nikfarjam et al. [135]	2015	8,800 blog sentences, 3,200 tweets	A,I,O (3)	H	CRF	Word embedding, CRF
	Jenhani et al. [34]	2016	80,000 tweets	A,O (2)	R, ODIN	Dictionary, Stanford CoreNLP	Rule-base,Feature-based, DT, SVM, LR, NB
	Liu et al. [138]	2016	1,800 blog sentences, 500 tweets	A,O (2)	H	MetaMap	Feature-based, Tree kernel-based, Ensemble method
<b>Semi-Supervised Learning</b>							

*Continued on next page*

Table 5.1 – Continued from previous page

Data Source	Literature	Year	Size	Label number	Labeling Method	NER	Method
EMR	Taewijit et al. [98]	2016	1.5M sentences	A,I (2)	DS	MetaMap	Distant supervision, OpenIE [87], Pattern-based
Literature	Kang et al. [139]	2014	1,644 abstracts	A,O (2)	H	Peregrine	Hierarchical graph-based, Shortest path
Social Media	Liu et al. [140]	2015	400 sentences	A,I,O (3)	H	Meta Map	Dependency tree, TSVM [141]
<b>Unsupervised Learning</b>							
EMR	Wang et al. [47]	2009	25,074 notes	none	none	MedLEE	Co-occurrence
Literature	Xu et al. [125]	2014	119M sentences	none	none	Parse tree	Pattern-based, Ranking
Social Media	Feldman et al. [134]	2015	0.1~1M messages	none	none	Dictionary, pattern	HPSG-based parser, post-processing of relation merging

**Labels:** A—ADR, I—IND, O—Other (ADR cause, ADR outcome, Non ADR, Negated ADR, Other)

**Labeling Method:** DS—Distant Supervision, H—Human, R—Rule-based

## 5.2 Problem Formulation

Firstly, the formal definition of distant supervision is presented then follow by problem formulation of multiple instance learning (MIL) concept.

### 5.2.1 Distant Supervision

Let  $\mathcal{K}$  denote knowledge bases regarding adverse drug reaction (ADR) and therapeutic indication (IND) that are obtained from SIDER<sup>1</sup> and DrugBank<sup>2</sup> databases,  $\mathcal{T}$  be a set of seeds, where  $\mathcal{T} \subseteq \mathcal{K}$ , and  $\mathcal{Y}$  is a set of labels, where  $\mathcal{Y} = \{ADR, IND\}$ , the data set of seeds  $\mathcal{T}$  in knowledge bases  $\mathcal{K}$  or an *entity-level* set can be defined as  $\mathcal{T} = \{(\mathbf{t}_1, y_1), (\mathbf{t}_2, y_2), \dots, (\mathbf{t}_N, y_N)\}$ , where  $\mathbf{t}_i = \{d_i, e_i\}$  is a seed,  $\mathbf{t}_i \in \mathcal{G}$  is 2-dimensional entities space which consists of a drug entity ( $d_i$ ) and an event entity ( $e_i$ ) that are defined in  $\mathcal{K}$ ,  $y_i \in \mathcal{Y}$  is a label corresponding  $\mathbf{t}_i$ , and  $N$  is a total number of seeds. Therefore, the data set of seeds can be derived as  $\mathcal{T} = \{(d_1, e_1, y_1), (d_2, e_2, y_2), \dots, (d_n, e_n, y_n)\}$ .

For instance, the drug *ramipril* associates with the adverse event *allergy* and the drug *ibuprofen* is used to treat the clinical event *arthritis* are supposed to exist in  $\mathcal{K}$ . A data set of seeds can be derived and used as a source of distant supervision such as  $\mathcal{T} = \{(ramipril_d, allergy_e, ADR), (ibuprofen_d, arthritis_e, IND)\}$ . These seeds are *entity-level* data that are used as knowledge for later processes.

Let  $\mathcal{C}$  be a clinical-records corpus from MIMIC<sup>3</sup>, which contains a set of discharge summary sentences  $\mathcal{S}$ . We transform each sentence into the three key elements i.e., a drug entity ( $d$ ), a key phrasal pattern entity ( $p$ ) and an event entity ( $e$ ), while semantic of such simplified texts is retrained. Given  $\mathbf{x}_j = \{d_j, p_j, e_j\}$  be a tuple obtained from an input sentence and  $\mathbf{x}_j \in \mathcal{H}$  is 3-dimensional entities space, in order to automatically generate labeled examples using distant supervision, the goal is to obtain a mapping function  $f : \mathcal{H} \rightarrow \mathcal{Y}$  that relates a drug-event pair of  $\{d_j, e_j\}$  to a relation label  $y_i$ , where  $(d_i, e_i, y_i)$  exists in  $\mathcal{T}$ ,  $d_j = d_i$ , and  $e_j = e_i$ . Finally, we can derive a set of labeled data  $\mathcal{D}_{\mathcal{L}} = \{(d_1, p_1, e_1, y_1), (d_2, p_2, e_2, y_2), \dots, (d_n, p_n, e_n, y_n)\}$ , namedly an *instance-level* data set, whereas  $n$  is a total number of mapped sentences.

---

<sup>1</sup><http://sideeffects.embl.de>

<sup>2</sup><https://www.drugbank.ca>

<sup>3</sup><https://mimic.physionet.org>

For example, a sentence “His ramipril were discontinued due to allergy and added to list in our medical records.” is supposed to exist in the corpus  $\mathcal{C}$ . Then the transformed sentence  $\mathbf{x}_1$  using a dependency tree can be simplified into the three key elements of a drug  $d_1 = \{\text{ramipril}_d\}$ , a key phrasal pattern  $p_1 = \{\text{be-discontinue-due-to}_p\}$  and an event  $e_1 = \{\text{allergy}_e\}$ , where a key phrasal pattern is applied in either the syntactically lemmatized lexicon or surface lexicon (e.g., was-discontinued-due-to), and can be employed as either word or phrase form (discuss later in section 5.3.2). From the mapping function  $f : \mathcal{H} \rightarrow \mathcal{Y}$ , we can project such sentence  $\mathbf{x}_1$  to a seed  $\{(\text{ramipril}_d, \text{allergy}_e, \text{ADR})\}$  in  $\mathcal{T}$  and transfer corresponding labels  $\text{ADR}$  to the sentence  $\mathbf{x}_1$ . Therefore, we can derive a labeled data by distant supervision as  $\{(\text{ramipril}_d, \text{be-discontinue-due-to}_p, \text{allergy}_e, \text{ADR})\} \in \mathcal{D}_{\mathcal{L}}$ . As another example, a sentence “The allergy improved despite ongoing treatment with ramipril.” is also supposed to exist in the corpus  $\mathcal{C}$ . The transformed sentence  $\mathbf{x}_2$  is  $\{\text{ramipril}_d, \text{improved-despite}_p, \text{allergy}_e\}$ . In the similar manner, we can use the mapping function  $f : \mathcal{H} \rightarrow \mathcal{Y}$  to assign the corresponding label of the entity pair  $\text{ramipril}_d$  and  $\text{allergy}_e$ . Therefore, the derived labeled data is  $\{\text{ramipril}_d, \text{improved-despite}_p, \text{allergy}_e, \text{ADR}\} \in \mathcal{D}_{\mathcal{L}}$ . However, the sentence  $\mathbf{x}_2$  might not express the correct clinical event of an adverse reaction. This is known as the noisy label and needs to handle by a particular technique such as MIL.

### 5.2.2 Multiple Instance Learning

In MIL concept, a bag- and an instance-level relation are equivalent to the entity- and the instance-level relation of drug-event relation derived by distant supervision respectively. Regarding the definition in section 2.5,  $\mathcal{X}$  be an instance space,  $\mathcal{Y}$  be a set of labels, where  $\mathcal{Y} = \{\text{ADR}, \text{IND}\}$ , the labeled data set  $\mathcal{D}_{\mathcal{L}}$  can be rewritten in the form of MIL as  $\mathcal{D}_{\mathcal{L}} = \{(B_1, y_1), (B_2, y_2), \dots, (B_n, y_n)\}$ , where  $B_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{im}\}$  is a set of multiple sentences that all sentences in a bag  $B_i$  corresponds to the same drug ( $d$ ) and event ( $e$ ),  $n$  is the number of bags, and  $m$  is the number of sentences in a bag and can be varied across a different bag. On the one hand, unlabeled instances ( $\mathcal{D}_{\mathcal{U}}$ ) are formed as a group of bags in the similar way but without a label as  $\mathcal{D}_{\mathcal{U}} = \{(B_1), (B_2), \dots, (B_n)\}$ . Our goal is both to train an instance classifier function

$f : \mathcal{X} \rightarrow \mathcal{Y}$  in the instance-space paradigm from  $\mathcal{D}_{\mathcal{L}}$  only (supervised learning) and attempt to infer the accurate label for each instance in the  $\mathcal{D}_{\mathcal{U}}$  set during the training process (transductive learning). The bag label, eventually, can be derived from an aggregation function of the instance level, and the model assessment is investigated through the model performance of the entity-level. Regarding noisy data labeling from distant supervision, the collective assumption and standard assumption with logical-OR aggregation for the bag label judgment are rather improper. The relaxed version of the MIL standard assumption is used in our proposed framework by assuming that the positive and negative bags are able to contain a mixture of either positive or negative instances, but the probability of *at least one* positive instance should be the maximum for the positive bag and vice versa. Consequently, to learn bag classifier  $f : 2^{\mathcal{X}} \rightarrow \mathcal{Y}$ , the estimated bag label from an instance classifier can be computed using Eq.5.1, where  $y_i$  is a label of a bag  $i$  (the entity-level label),  $y_{ij}$  is a label of the instance-level and possibly different for each sentence instance  $j$  within the same bag  $i$ , and  $n$  is the total number of sentences in the bag.

$$p(y_i|B_i) = \max_{j \in \{1, \dots, n\}} p(y_{ij}|\mathbf{x}_{ij}) \quad (5.1)$$

Generally, the training data are not sufficient for parameters training. In order to learn such classifier function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , we make use of the iterative EM technique with transductive learning setting to estimate the posterior probability  $p(y|\mathbf{x})$  through the two parameters, i.e., prior probability  $p(y)$  and class-conditional density  $p(\mathbf{x}|y)$ , of the generative model.

### 5.3 Proposed Method

In this section, a framework for ADR identification from the clinical text is proposed in order to overcome the shortcomings of the existing research; (i) the lack of domain experts for labeling examples, and (ii) intractable processing of a large-scale unstructured clinical text. Figure 5.1 illustrates the overview of the proposed method. The proposed adverse drug reaction identification framework consists of the three main tasks.

1. In the relation generation, drug-event pairs  $(d, e)$  's are extracted from a corpus together with their patterns  $(p)$  using sentence boundary detection (see Section 3.2.2), named entity recognition (see Section 3.2.3) and parsing.
2. In the automatic data labeling (see Section 5.3.1), distant supervision assigns a relation label  $(y)$  to each drug-event pair  $(d, e)$  obtained from the relation generation with its pattern  $(p)$  if such relation exists in knowledge base. The *Silver-Standard* data set is labeled data in the experiment. Here, two types of output datasets are a set of labeled data  $(\mathcal{D}_L)$ , composed of  $(d, p, e, y)$  's extracted from a MIMIC-III corpus, where the labels  $(y)$  are defined for the drug-event pairs  $(d, e)$  in the knowledge base, and a set of unlabeled data  $(\mathcal{D}_U)$ , composed of  $(d, p, e)$  's extracted from a corpus, where the labels do not exist for the drug-event pairs  $(d, e)$  in the knowledge base.
3. In this relation classification, this work proposes three types of generative models with independent/dependent expectation-maximization (EM) model (iEM/dEM); (i) transductive learning with iEM (baseline), (ii) supervised learning with dEM, and (iii) transductive learning with dEM.

The former issue is relevant to assign a label to unlabeled instances for training examples preparation, so-called *data labeling*. The distant supervision paradigm replaces human efforts by employing the heuristics or rules through the fact from the knowledge bases. The two sources of knowledge bases from SIDER and DrugBank are utilized in this work, while, ADR and IND labels are considered as classification outputs. For the reason that ADR and IND are used as the target labels because the projection from a finite set of known facts such as ADR or IND is easier and more accurate than the consideration from an infinite set of hidden knowledge such as non-ADR. While distant supervision can supply a label to an unlabeled instance by simply looking up from knowledge bases, the labeled data set by this method is formed as MIL problem which training labels are associated with sets of instance examples rather than individual examples.

As for the latter issue, applying phrase-based method and dependency representation may improve the model performance. The main idea is that a sentence re-

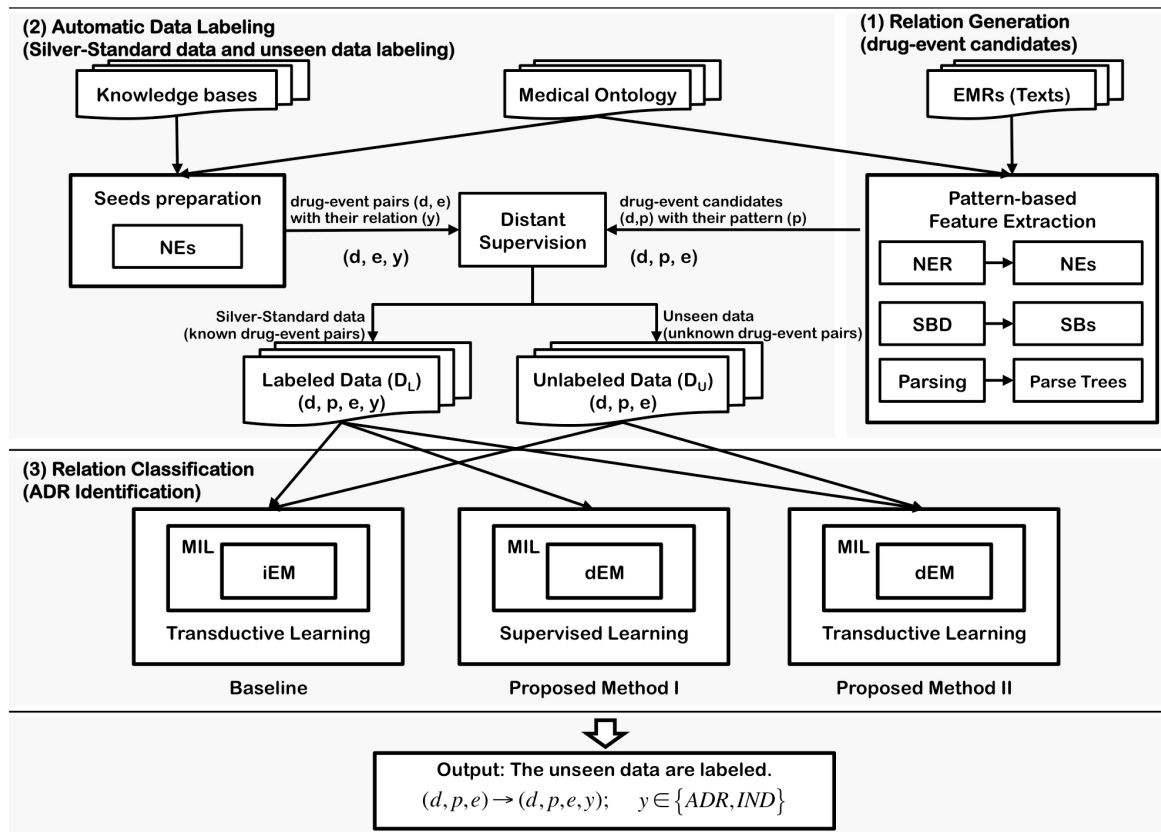


Fig. 5.1: Overview of the proposed adverse drug reaction identification framework

garding harmful from an adverse reaction or beneficial treatment can be simplified into the three key elements, *a drug*, *a key phrasal pattern*, and *an event*; and the dependency among such three elements has significance. Such key phrasal pattern implies a semantic relation between any pair of drug and event entities. A key phrasal pattern-based method for ADR identification in Chapter 4 is employed. The method exhibits the high precision; notwithstanding its drawback is low recall rate due to a limit to the number of key phrasal patterns and the utilization of simple models. In this chapter, such key phrasal pattern-based method is extended with more sophisticated models, which is expected to be able to retain the high precision and improve retrieval performance. The Expectation Maximization (EM), an iterative method, is incorporated with Markov property assumption to draw conditional probability distribution of pattern-based feature (dEM). Finally, unlabeled data is leveraged through the transductive inference as semi-supervised learning to enhance the performance of the proposed framework. For performance evaluation, EM with the independence assump-

tion is constructed through naïve bayes (iEM) as the baseline. The proposed methods is compared to multiple advanced methods; Multiple-Instance Support Vector Machine (MISVM), Multiple-Instance Naïve Bayes (MINB), Multiple-Instance Logistic Regression (MILR) and Transductive Support Vector Machine (TSVM). The multiple numbers of parameters such as pattern types, pattern-weighting models, initial and iterative weightings relation labels for unlabeled data are investigated throughout three alternative MIL models; iEM with transductive inference setting (baseline), dEM supervised learning and dEM with transductive inference.

### 5.3.1 Distantly Supervised Ground Truth

Figure 5.2 displays information extraction from sentences in the MIMIC corpus, with the output of drug-key phrasal pattern-event tuples as candidates of ADR or IND relation. This process involves named entity recognition, sentence boundary detection and parsing. On the left-hand side, the first block, UMLS is made use to discover particular two semantic types; *drug* and *symptom*. Concept Unique Identifiers (CUI) is taken place on the discovered terms for the generalization (the second block). The fourth block, the dependency parse tree derived from OpenIE contributes to the relation tuples extraction. The simplified text, relation tuple-base text encapsulation, is considering a pair of entities (rectangle and triangle represent drug and symptom entities respectively) and a phrase-pattern (pentagon presents the pattern) which is implied the semantic relation between the pair. Here, the MetaMap [81] is used for named entity recognition, the developed in-house program for sentence boundary detection, and Stanford CoreNLP’s OpenIE for parsing. After extracting relation candidate tuples  $(arg_1, pattern, arg_2)$ , only the tuples that include drug name and event name as  $arg_1$  and  $arg_2$  or vice versa are selected. The output is in the form of (a drug, a key phrasal pattern, an event).

The automatic labeling process using distant supervision is illustrated in Figure 5.3. Block-1 expresses the data labeling using the fact from external sources (KB seeds). The  $\mathcal{D}_{\mathcal{L}}$  is a data set that a pair of drug and event entities can be mapped to a set of KB seeds through the distant supervision. Hence, all sentences that correspond to a

same drug-event pair are assigned to the same bag and a same label (labeled data  $\mathcal{D}_{\mathcal{L}}$ ) regarding a label of such drug-event pair in a set of seeds from knowledge base. Finally, such  $\mathcal{D}_{\mathcal{L}}$  set is used as a training data. However, to reduce the ambiguity of the ground truth from knowledge base supervision, a pair of  $(d, e)$  that is found to exhibit both *ADR* and *IND* semantic relations is excluded. Given a set of sentences  $\mathcal{X}$ , the training set  $\mathcal{D}_{\mathcal{L}}$  is in the form  $\{(B_1, y_1), (B_2, y_2), \dots, (B_n, y_n)\}$ , where  $B_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{im}\}$ . In the Block-1 of Fig. 5.3, the first bag ( $Bag_1$ ) consists of two sentences that correspond to the same entity-level of drug  $d_1$  and event  $e_1$ . The second bag ( $Bag_2$ ) contains only one sentence relevant to drug  $d_2$  and event  $e_4$ .

Finally, all sentences that are able to be assigned a label by distant supervision are referred as the set of *labeled data*  $\mathcal{D}_{\mathcal{L}}$  and the remaining data that are not matched by distant supervision is used as *unlabeled data*  $\mathcal{D}_{\mathcal{U}}$ .

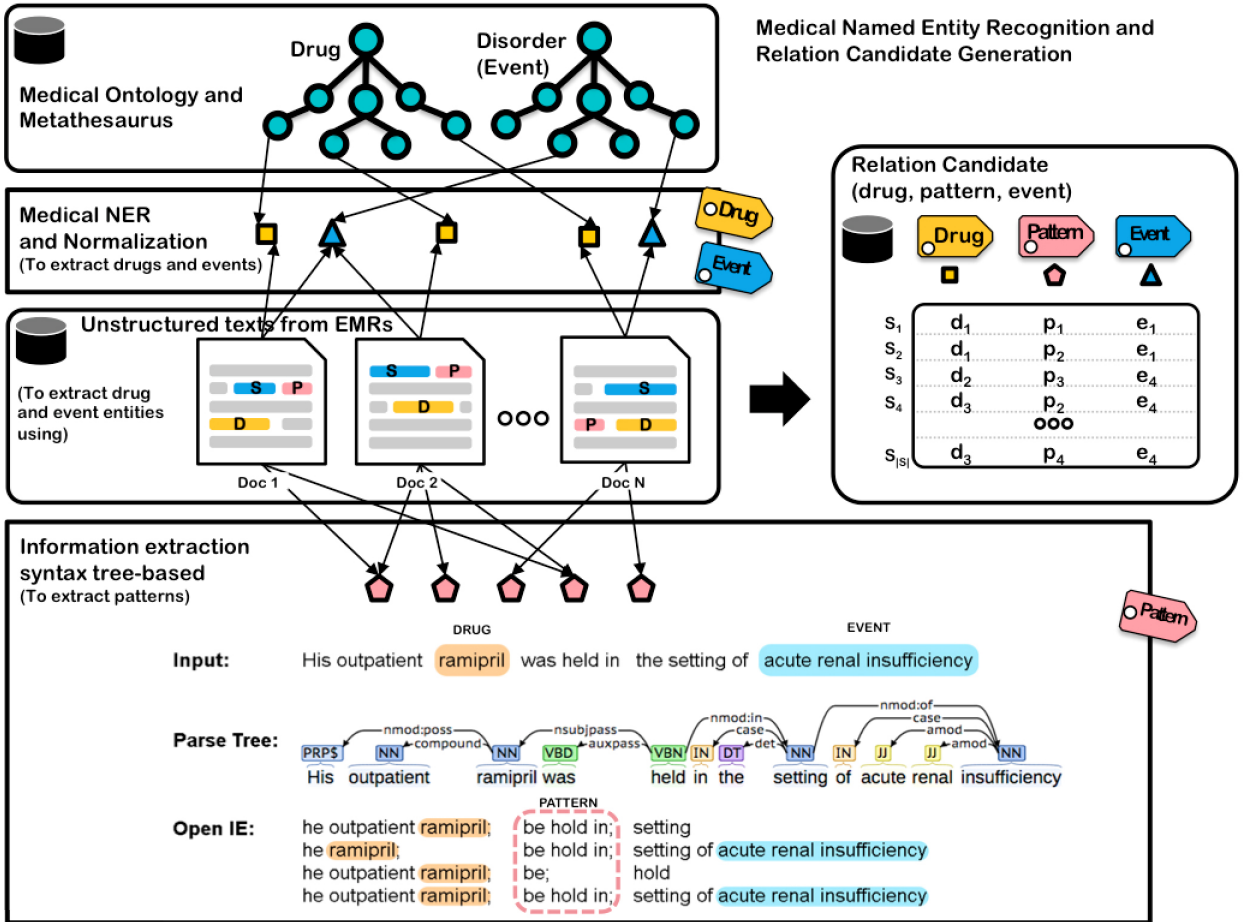


Fig. 5.2: Medical named entity recognition and relation candidate generation.

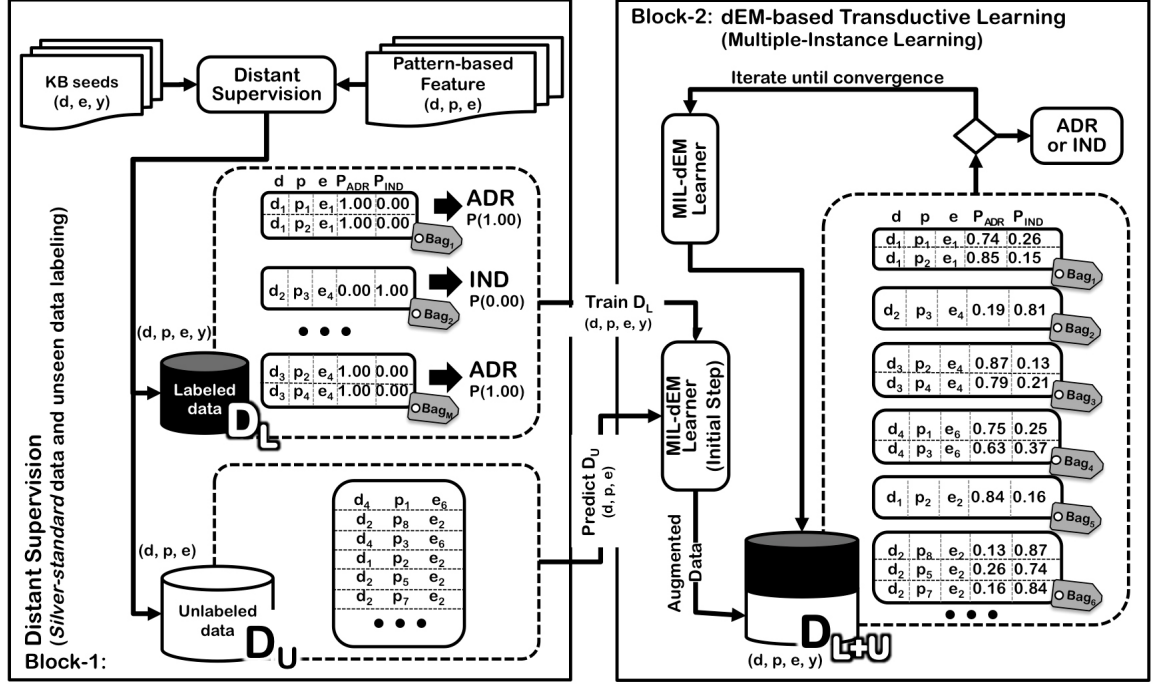


Fig. 5.3: Block-1 distant supervise for automatic data labeling. Block-2 depicts the proposed MIL-dEM method.

### 5.3.2 Document Representation

The six pattern types as feature extraction across the three alternative pattern weighting models as text representation are assessed. The Bag-of-Words (BOW) feature extraction with iEM model is used as the based line.

#### Feature Extraction for Medical Textual Data

To recognize a relationship between a drug and an event, our approach generates a set of relation candidates (drug-event pairs) from medical records in the form of (drug, pattern, event). Table 5.2 depicts examples of multiple types of feature extraction and drug-event candidates. Our work considers three parameters related to representing such relation candidates. The first parameter, called relation boundary constraint, defines the potential of using surrounding context for determining drug-event relations while the second and third parameters, called syntactic lemmatization and pattern granularity constraints, are related to patterns used to detect drug-event relations, as follows.

- (i) *boundary constraint*: considering of surrounding contexts of relation tuples regarding window size  $w$  (context base  $C$ ). In this work,  $w$  is a size of a sentence. Another one is considering only relation tuples of a sentence without surrounding context (no context  $N$ )
- (ii) *syntactic lemmatization*: considering a canonical form of a term that appears in a pattern within relation tuples (surface lexicon  $S$ ) or considering a dictionary form of the text (syntactically lemmatized lexicon  $L$ ).
- (iii) *pattern granularity*: treating a pattern in relation tuples as an individual term (word form  $W$ ) or a group of words (phrase form  $P$ ).

### Pattern-weighting models

- (i) *Bernoulli (binary) document model ( $B$ )*: A document (hereinafter referred to as a sentence denoted by  $x$ ) can be represented in the form of a vector each element of which corresponds to a term (i.e., word, phrase) denoted by  $w$  with a value of either 0 or 1 for presence or absence of such term, respectively.

$$\mathbf{x}_B = \{B(x, w_1), B(x, w_2), \dots, B(x, w_{|W|})\}, \quad (5.2)$$

where  $\mathbf{x}_B$  presents a sentence  $x$  in the form of a binary vector,  $B(x, w_i) = 1$  when  $w_i$  is the  $i$ -th term in the sentence  $x$  (otherwise 0), and  $w_i$  is a term in the universe  $W$ .

- (ii) *Multinomial (frequency) document model*: A sentence is expressed by a vector of term frequency ( $TF$ ) as

$$\mathbf{x}_{TF} = \{TF(x, w_1), TF(x, w_2), \dots, TF(x, w_{|W|})\}; TF(x, w_i) = \frac{f_x(w_i)}{|x|}, \quad (5.3)$$

where  $\mathbf{x}_{TF}$  is a sentence  $x$  in the form of a  $TF$  vector,  $TF(x, w_i)$  expresses the normalized frequency of the  $i$ -th term  $w_i$  by the sentence size  $|x|$ , and  $f_x(w_i)$  is the frequency that the term  $w_i$  occurs in the sentence  $x$ . As another option, a document can also be expressed by a vector of term frequency-inverse document

Table 5.2: Types of feature extraction for a given sentence.

Sentences	Types	Example of feature representation
On arrival here, <u>propofol</u> <i>was held due to hypotension</i> .	<i>N-L-P</i>	C0033487 <i>be-held-due-to</i> C0020649
	<i>N-L-W</i>	C0033487 <i>be held due to</i> C0020649
	<i>N-S-P</i>	C0033487 <i>was-held-due-to</i> C0020649
	<i>N-S-W</i>	C0033487 <i>was held due to</i> C0020649
	<i>C-L-P</i>	On arrival here, C0033487 <i>be-held-due-to</i> C0020649
	<i>C-L-W</i>	On arrival here, C0033487 <i>be held due to</i> C0020649
	<i>C-S-P</i>	On arrival here, C0033487 <i>was-held-due-to</i> C0020649
	<i>C-S-W</i>	On arrival here, C0033487 <i>was held due to</i> C0020649
	<i>BOW</i>	On arrival here, propofol was held due to hypotension
<u>Phenylephrine</u> drip <i>was started for hypotension</i> .	<i>N-L-P</i>	C0031469 <i>be-started-for</i> C0020649
	<i>N-L-W</i>	C0031469 <i>be start for</i> C0020649
	<i>N-S-P</i>	C0031469 <i>was-started-for</i> C0020649
	<i>N-S-W</i>	C0031469 <i>was started for</i> C0020649
	<i>C-L-P</i>	C0031469 drip <i>be-started-for</i> C0020649
	<i>C-L-W</i>	C0031469 drip <i>be start for</i> C0020649
	<i>C-S-P</i>	C0031469 drip <i>was-started-for</i> C0020649
	<i>C-S-W</i>	C0031469 drip <i>was started for</i> C0020649
	<i>BOW</i>	Phenylephrine drip was started for hypotension

The first character of Types: *N*–No context, *C*–Context

The second character of Types: *L*–Syntactically lemmatized lexicon, *S*–Surface lexicon

The third character of Types: *W*–Word, *P*–Phrase

BOW stands for Bag-Of-Words.

CUI C0033487 is a UMLS concept of propofol, CUI C0031469 is a UMLS concept of phenylephrine and CUI C0020649 is a UMLS concept of hypotension.

frequency ( $TFIDF$ ) as

$$\begin{aligned}\mathbf{x}_{TFIDF} &= \{TF(x, w_1) \cdot IDF(w_1), TF(x, w_2) \cdot IDF(w_2), \dots, TF(x, w_{|W|}) \cdot IDF(w_{|W|})\}; \\ TF(x, w_i) &= \frac{f_x(w_i)}{|x|}; \\ IDF(w_i) &= \log \frac{|\mathcal{X}|}{|\{x|x \in \mathcal{X} \text{ and } B(x, w_i) = 1\}|},\end{aligned}\tag{5.4}$$

where  $\mathbf{x}_{TFIDF}$  is a sentence  $x \in \mathcal{X}$  (the document universe), in the form of a  $TFIDF$  vector,  $IDF(w_i)$  expresses the inverse document frequency, corresponding to the logarithm of the ratio of the total number of sentences in the universe  $|\mathcal{X}|$  to the number of sentences that contain the  $i$ -th term  $w_i$ .

### 5.3.3 Transductive Learning for Relation Extraction

This section describes two EM-based probabilistic classification models; one with independent assumption ( $iEM$ ) and the other with dependent representation assumption ( $dEM$ ).

#### EM model with Naïve Bayes independent assumption (iEM)

Let  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{|\mathcal{X}|}\}$  be a set of sentences,  $\mathbf{x}_i = \{w_{i1}, w_{i2}, \dots, w_{i|\mathcal{X}_i|}\}$  is a sentence that includes  $|\mathcal{X}_i|$  terms, and  $C = \{c_1, c_2, \dots, c_{|C|}\}$  be the set of possible classes. The probability that the sentence  $\mathbf{x}_i$  has  $c_k$  as its class ( $y_i = c_k$ ) can be formulated as shown in Eq.5.5.

$$\begin{aligned}p(y_i = c_k | \mathbf{x}_i) &= \frac{p(c_k) p(\mathbf{x}_i | c_k)}{p(\mathbf{x}_i)} \\ &= \frac{p(c_k) p(\mathbf{x}_i | c_k)}{\sum_{k=1}^{|C|} p(c_k) p(\mathbf{x}_i | c_k)}\end{aligned}\tag{5.5}$$

While in most situations, it is possible to obtain the class  $p(c_k)$  simply from the training set, the generative probability of  $\mathbf{x}_i$  given a class  $c_k$  usually suffer from insufficient training data. As done by several works, the assumption of independence, usually called Naïve Bayes (NB), can be applied to alleviate this sparseness problem as expressed in Eq.5.6.

$$\begin{aligned}
p(\mathbf{x}_i|c_k) &= p(w_{i1}, w_{i2}, \dots, w_{i|\mathbf{x}_i|}|c_k) \\
&= p(w_{i1}|c_k) \cdot p(w_{i2}|w_{i1}, c_k) \cdot \dots \cdot p(w_{i|\mathbf{x}_i|}|w_{i1}, w_{i2}, \dots, w_{i(|\mathbf{x}_i|-1)}, c_k) \\
&\approx p(w_{i1}|c_k) \cdot p(w_{i2}|c_k) \cdot \dots \cdot p(w_{i|\mathbf{x}_i|}|c_k) \\
&= \prod_{j=1}^{|\mathbf{x}_i|} p(w_{ij}|c_k)
\end{aligned} \tag{5.6}$$

Therefore, the NB text classifier can be rewritten in the form as shown in Eq.5.7.

$$p(y_i = c_k|\mathbf{x}_i) = \frac{p(c_k) \prod_{j=1}^{|\mathbf{x}_i|} p(w_{ij}|c_k)}{\sum_{k=1}^{|C|} p(c_k) \prod_{j=1}^{|\mathbf{x}_i|} p(w_{ij}|c_k)} \tag{5.7}$$

Here, it is necessary to estimate two sets of parameters, denoted by  $\theta$ , of Expectation-Maximization (EM) algorithm. The first parameter set is the class-conditional probability of any term  $w_q \in W$  given the class  $c_k$  while the other one is the probability of the class  $c_k$ . The parameter set is defined by Eq.5.8.

$$\theta = \{p^{(t+1)}(w_q|c_k), p^{(t+1)}(c_k)\} \tag{5.8}$$

In the expectation step (E-step), for each iteration the  $\theta$  parameter of the previous step is applied to re-estimate the model probability as shown in Eq.5.9. In our experiment, the convergence threshold is  $10^{-7}$  and the maximum number of iterations is set to 50.

$$p^{(t)}(y_i = c_k|\mathbf{x}_i) = \frac{p^{(t-1)}(c_k) \prod_{j=1}^{|\mathbf{x}_i|} p^{(t-1)}(w_{ij}|c_k)}{\sum_{k=1}^{|C|} p^{(t-1)}(c_k) \prod_{j=1}^{|\mathbf{x}_i|} p^{(t-1)}(w_{ij}|c_k)} \tag{5.9}$$

For the maximization step (M-step), with a Laplace smoothing factor  $\lambda > 0$ , the  $(t+1)$ -th-iteration probability of  $p^{(t+1)}(w_q|c_k)$  and  $p^{(t+1)}(c_k)$  can be estimated from the  $t$ -th-iteration probability. The maximum likelihood estimation for NB is simply computed from an empirical corpus using Eq.5.10 and Eq.5.11.

$$p^{(t+1)}(w_q|c_k) = \frac{\lambda + \sum_{i=1}^{|\mathbf{X}|} N(w_q, \mathbf{x}_i) p^{(t)}(y_i = c_k|\mathbf{x}_i)}{\lambda|W| + \sum_{r=1}^{|W|} \sum_{i=1}^{|\mathbf{X}|} N(w_r, \mathbf{x}_i) p^{(t)}(y_i = c_k|\mathbf{x}_i)}, \tag{5.10}$$

whereas  $W$  is a total number of terms and any term  $w_z \in W$

$$p^{(t+1)}(c_k) = \frac{\lambda + \sum_{i=1}^{|\mathbf{X}|} p^{(t)}(y_i = c_k|\mathbf{x}_i)}{\lambda|C| + |\mathbf{X}|} \tag{5.11}$$

The following demonstrates an example of the applying above formulations with the key phrasal pattern-based feature. Given the  $L$ - $P$  feature representation of  $\mathbf{x}_i =$

(*C0033487, be-hold-due-to, C0020649*) corresponds to relation tuple  $(d_i, p_i, e_i)$  obtained from an input sentence where the pattern  $p_i$  be the phrase form, we can estimate  $p(y_i = c_k | \mathbf{x}_i)$  as expressed in Eq.5.12.

$$p^{(t)}(y_i = c_k | \mathbf{x}_i) = \frac{p^{(t-1)}(c_k) \cdot p^{(t-1)}(C0033487|c_k) \cdot p^{(t-1)}(be-hold-due-to|c_k) \cdot p^{(t-1)}(C0020649|c_k)}{\sum_{k=1}^{|C|} p^{(t-1)}(c_k) \cdot p^{(t-1)}(C0033487|c_k) \cdot p^{(t-1)}(be-hold-due-to|c_k) \cdot p^{(t-1)}(C0020649|c_k)} \quad (5.12)$$

Another example, given the  $L-W$  feature representation of the same sentence  $\mathbf{x}_i = \{C0033487, be, hold, due, to, C0020649\}$  corresponds to relation tuple  $(d_i, p_i, e_i)$  where the pattern  $p_i$  is in the word form. We can compute the class probability of the given texts  $p(y_i = c_k | \mathbf{x}_i)$  as shown in Eq.5.13.

$$p^{(t)}(y_i = c_k | \mathbf{x}_i) = \frac{p^{(t-1)}(c_k) \cdot p^{(t-1)}(C0033487|c_k) \cdot p^{(t-1)}(be|c_k) \cdot p^{(t-1)}(hold|c_k) \cdot p^{(t-1)}(due|c_k) \cdot p^{(t-1)}(to|c_k) \cdot p^{(t-1)}(C0020649|c_k)}{\sum_{k=1}^{|C|} p^{(t-1)}(c_k) \cdot p^{(t-1)}(C0033487|c_k) \cdot p^{(t-1)}(be|c_k) \cdot p^{(t-1)}(hold|c_k) \cdot p^{(t-1)}(due|c_k) \cdot p^{(t-1)}(to|c_k) \cdot p^{(t-1)}(C0020649|c_k)} \quad (5.13)$$

### EM model with dependency representation (dEM)

We introduce a dependency representation as an alternative model representation that is based on the same intuitions as the NB model but less restriction regarding the implicitly strong independence assumptions. This dependency representation is an efficient factorization of the joint probability distributions over a set of three random variables  $w_q, w_r$  and  $w_s$ , where each variable is a domain of possible values, i.e., drug, key phrasal pattern, event. We extend the dependency representation with iterative learning by EM approach in order to align the model assumption to the natural language and also figure out an unseen random variable using probability estimation based on an existing prior knowledge. This dependency representation is also known as Bayesian Networks (BN) and the conditional probability of independent variable given a class probability can be derived by the chain rule as shown in Eq.5.14.

$$\begin{aligned} p(\mathbf{x}_i | c_k) &= p(w_{iq}, w_{ir}, w_{is} | c_k) \\ &= p(w_{iq} | c_k) \cdot p(w_{ir} | w_{iq}, c_k) \cdot p(w_{is} | w_{iq}, w_{ir}, c_k) \end{aligned} \quad (5.14)$$

Therefore, the BN text classifier can be rewritten in the form as shown in Eq.5.15.

$$p(y_i = c_k | \mathbf{x}_i) = \frac{p(c_k) \cdot p(w_{iq} | c_k) \cdot p(w_{ir} | w_{iq}, c_k) \cdot p(w_{is} | w_{iq}, w_{ir}, c_k)}{\sum_{k=1}^{|C|} p(c_k) \cdot p(w_{iq} | c_k) \cdot p(w_{ir} | w_{iq}, c_k) \cdot p(w_{is} | w_{iq}, w_{ir}, c_k)} \quad (5.15)$$

According to the core of BN representation, a random variable is represented by a node in a directed acyclic graph, and an edge between any two nodes is presented by an arrow line which implies a direct influence of one node on another node. Given a sentence  $\mathbf{x}_i$  with three elements  $(w_{iq}, w_{ir}, w_{is})$  in the form of a relation tuple  $(d_i, p_i, e_i)$ , there are three factorized ways (3!) as alternative model skeletons of the dependency representation through the chain rule. We, hence, propose the linear interpolation in order to weight and combine the probability estimation from all of the possible dependency representations as defined by Eq.5.16.

$$\begin{aligned}
p(\mathbf{x}_i|c_k) &= p(w_{iq}, w_{ir}, w_{is}|c_k) \\
&\approx \gamma_1 [p(w_{iq}|c_k) \cdot p(w_{ir}|w_{iq}, c_k) \cdot p(w_{is}|w_{iq}, w_{ir}, c_k)] \\
&\quad + \gamma_2 [p(w_{iq}|c_k) \cdot p(w_{is}|w_{iq}, c_k) \cdot p(w_{ir}|w_{iq}, w_{is}, c_k)] \\
&\quad + \gamma_3 [p(w_{ir}|c_k) \cdot p(w_{iq}|w_{ir}, c_k) \cdot p(w_{is}|w_{iq}, w_{ir}, c_k)] \\
&\quad + \gamma_4 [p(w_{ir}|c_k) \cdot p(w_{is}|w_{ir}, c_k) \cdot p(w_{iq}|w_{ir}, w_{is}, c_k)] \\
&\quad + \gamma_5 [p(w_{is}|c_k) \cdot p(w_{iq}|w_{is}, c_k) \cdot p(w_{ir}|w_{iq}, w_{is}, c_k)] \\
&\quad + \gamma_6 [p(w_{is}|c_k) \cdot p(w_{ir}|w_{is}, c_k) \cdot p(w_{iq}|w_{ir}, w_{is}, c_k)],
\end{aligned}$$

such that the total  $\gamma$  is 1;  $\sum_{i=1}^6 \gamma_i = 1$

(5.16)

Generally, the linear interpolation method of three random variables can be estimated from the combination of two random variables and an individual random variable. Similarly, two random variables are able to approximate from an individual random variable as well. For instance, given two history terms  $w_{iq}$  and  $w_{ir}$  in a sentence  $\mathbf{x}_i$ , the interpolation is comparatively estimated from an individual random variable and two random variables as shown in Eq.5.17.

$$\begin{aligned}
p(w_{ir}|w_{iq}, c_k) &= \beta_1 p(w_{ir}|c_k) + \beta_2 p(w_{ir}|w_{iq}, c_k),
\end{aligned}$$

such that the total  $\beta$  is 1;  $\sum_{i=1}^2 \beta_i = 1$

(5.17)

Another instance, three history terms  $(w_{iq}, w_{ir}, w_{is})$  in a sentence  $\mathbf{x}_i$  are given, the likelihood estimation as shown in Eq.5.18 can be derived similarly as the previous estimator by interpolation of an individual random variable, two random variables and three random variables estimators.

$$p(w_{is}|w_{iq}, w_{ir}, c_k) = \alpha_1 p(w_{is}|c_k) + \alpha_2 p(w_{is}|w_{iq}, c_k) + \alpha_3 p(w_{is}|w_{ir}, c_k) + \alpha_4 p(w_{is}|w_{iq}, w_{ir}, c_k),$$

such that the total  $\alpha$  is 1;  $\sum_{i=1}^4 \alpha_i = 1$

(5.18)

Finally, we compute  $p(w_{iq}|w_{ir}, c_k)$ ,  $p(w_{iq}|w_{is}, c_k)$ ,  $p(w_{ir}|w_{is}, c_k)$ ,  $p(w_{is}|w_{iq}, c_k)$  and  $p(w_{is}|w_{ir}, c_k)$  with the similar manner as Eq.5.17, and calculate  $p(w_{iq}|w_{ir}, w_{is}, c_k)$  and  $p(w_{ir}|w_{iq}, w_{is}, c_k)$  using the same way as shown in Eq.5.18.

In the same manner as the naïve bayese model, it is necessary to estimate the four sets of parameters  $\theta$  as shown in Eq.5.19 whereas any terms  $w_q, w_r, w_s \in W$ .

$$\theta = \{p^{(t+1)}(w_q|c_k), p^{(t+1)}(w_q|w_r, c_k), p^{(t+1)}(w_q|w_r w_s, c_k), p^{(t+1)}(c_k)\}$$
(5.19)

The iterative learning using EM approach is applied to estimate the parameters  $\theta$ . For the E-step, for each iteration the  $\theta$  parameter is applied to re-estimate the model probability as shown in Eq.5.20 and Eq.5.21. This process will repeat until convergence. The same setting as iEM model, the value of  $10^{-7}$  for the convergence threshold and the value of 50 for the maximum number of iterations, is applied for dEM model as well.

$$\begin{aligned} p^{(t-1)}(\mathbf{x}_i|c_k) &\approx \gamma_1 [p^{(t-1)}(w_{iq}|c_k) \cdot p^{(t-1)}(w_{ir}|w_{iq}, c_k) \cdot p^{(t-1)}(w_{is}|w_{iq}, w_{ir}, c_k)] \\ &+ \gamma_2 [p^{(t-1)}(w_{iq}|c_k) \cdot p^{(t-1)}(w_{is}|w_{iq}, c_k) \cdot p^{(t-1)}(w_{ir}|w_{iq}, w_{is}, c_k)] \\ &+ \gamma_3 [p^{(t-1)}(w_{ir}|c_k) \cdot p^{(t-1)}(w_{iq}|w_{ir}, c_k) \cdot p^{(t-1)}(w_{is}|w_{iq}, w_{ir}, c_k)] \\ &+ \gamma_4 [p^{(t-1)}(w_{ir}|c_k) \cdot p^{(t-1)}(w_{is}|w_{ir}, c_k) \cdot p^{(t-1)}(w_{iq}|w_{ir}, w_{is}, c_k)] \\ &+ \gamma_5 [p^{(t-1)}(w_{is}|c_k) \cdot p^{(t-1)}(w_{iq}|w_{is}, c_k) \cdot p^{(t-1)}(w_{ir}|w_{iq}, w_{is}, c_k)] \\ &+ \gamma_6 [p^{(t-1)}(w_{is}|c_k) \cdot p^{(t-1)}(w_{ir}|w_{is}, c_k) \cdot p^{(t-1)}(w_{iq}|w_{ir}, w_{is}, c_k)] \end{aligned}$$
(5.20)

$$p^{(t)}(y_i = c_k|\mathbf{x}_i) = \frac{p^{(t-1)}(c_k) \cdot p^{(t-1)}(\mathbf{x}_i|c_k)}{\sum_{k=1}^{|C|} p^{(t-1)}(c_k) \cdot p^{(t-1)}(\mathbf{x}_i|c_k)}$$
(5.21)

For the M-step, the Laplace smoothing factor  $\lambda > 0$  is implemented as well as in NB model to avoid zero count issue. However, with the BN dependency representation, there are four parameters estimation of the  $(t + 1)$ -th-iteration probability of

$p^{(t+1)}(w_q|w_r, w_s, c_k)$ ,  $p^{(t+1)}(w_q|w_r, c_k)$ ,  $p^{(t+1)}(w_q|c_k)$  and  $p^{(t+1)}(c_k)$ , which can be estimated from  $t$ -th-iteration probability as expressed in Eq.5.22-5.25.

$$p^{(t+1)}(w_q|c_k) = \frac{\lambda + \sum_{i=1}^{|X|} N(w_q, \mathbf{x}_i) p^{(t)}(y_i = c_k | \mathbf{x}_i)}{\lambda |W| + \sum_{z=1}^{|W|} \sum_{i=1}^{|X|} N(w_z, \mathbf{x}_i) p^{(t)}(y_i = c_k | \mathbf{x}_i)}, \quad (5.22)$$

$$p^{(t+1)}(w_q|w_r, c_k) = \frac{\lambda + \sum_{i=1}^{|X|} N(w_q, \mathbf{x}_i) p^{(t)}(y_i = c_k | w_r, \mathbf{x}_i)}{\lambda |W| + \sum_{z=1}^{|W|} \sum_{i=1}^{|X|} N(w_z, \mathbf{x}_i) p^{(t)}(y_i = c_k | w_r, \mathbf{x}_i)}, \quad (5.23)$$

$$p^{(t+1)}(w_q|w_r, w_s, c_k) = \frac{\lambda + \sum_{i=1}^{|X|} N(w_q, \mathbf{x}_i) p^{(t)}(y_i = c_k | w_r, w_s, \mathbf{x}_i)}{\lambda |W| + \sum_{z=1}^{|W|} \sum_{i=1}^{|X|} N(w_z, \mathbf{x}_i) p^{(t)}(y_i = c_k | w_r, w_s, \mathbf{x}_i)}, \quad (5.24)$$

whereas  $W$  is a total number of terms and any term  $w_z \in W$

$$p^{(t+1)}(c_k) = \frac{\lambda + \sum_{i=1}^{|X|} p^{(t)}(y_i = c_k | \mathbf{x}_i)}{\lambda |C| + |X|} \quad (5.25)$$

Then, we can derive  $p^{(t+1)}(w_r|c_k)$  and  $p^{(t+1)}(w_s|c_k)$  using the similar calculation as Eq.5.22. For the dependency representations of two random variables  $w$ , i.e.,  $p^{(t+1)}(w_q|w_s, c_k)$ ,  $p^{(t+1)}(w_r|w_q, c_k)$ ,  $p^{(t+1)}(w_r|w_s, c_k)$ ,  $p^{(t+1)}(w_s|w_q, c_k)$  and  $p^{(t+1)}(w_s|w_r, c_k)$  can be computed by following the similar approach as Eq.5.23. Similarly, the estimation of  $p^{(t+1)}(w_r|w_q, w_s, c_k)$  and  $p^{(t+1)}(w_s|w_q, w_r, c_k)$  can be obtained by the same way as shown in Eq.5.24. Finally, the coefficients  $\gamma, \beta, \alpha$  of interpolation approach are employed in order to weight the knowledge from multiple dependency representations. The algorithm 2 explains pseudo-code for iEM model and algorithm 3 expresses the proposed dEM method.

### 5.3.4 The Incorporation of Unlabeled Data

In the environment of insufficient labeled data, semi-supervised learning is one solution that utilizes an inexpensive and ubiquitous source of data. The transductive learning [142], one type of semi-supervised learning, begins its process with the make use of a limited number of labeled data ( $\mathcal{D}_L$ ) to build a rough model and then aggregated a large number of unlabeled data ( $\mathcal{D}_U$ ) (test set) to revise and improve the model iteratively. In the experiment, we investigated the three alternative approaches of initialization and iterative weighting of relation labels for unlabeled data incorporation.

---

**Algorithm 2:** Pseudo-code for EM with NB independent assumption (iEM)

---

**Input:**  
 $|C|$  = the number of labels  
 $T$  = the maximum number of iteration  
**Output:**  $\theta$  parameter

```

1  $t \leftarrow 0$ 
2  $\theta = \{p^{(t+1)}(w_q, c_k), p^{(t+1)}(c_k)\}; \sum_{k=1}^{|C|} p^{(t+1)}(c_k) = 1$ 
3 repeat
4   for  $i = 1$  to  $n$  do
5     E-step:
6       Estimate model probability :  $p^{(t)}(y_i = c_k | \mathbf{x}_i)$  (Eq.5.9)
7     M-step:
8       Update class-conditional probability :  $p^{(t+1)}(w_q | c_k)$  (Eq.5.10)
9       Update class probability :  $p^{(t+1)}(c_k)$  (Eq.5.11)
10     $t \leftarrow t + 1$ 
11 until convergence or  $t = T$ 

```

---

- (i)  $T_{p_{ML}}$ : This method is equivalent to the general transductive learning, in which the label of the test set  $\mathcal{D}_{\mathcal{U}}$  can be derived by a classifier that is trained on the  $\mathcal{D}_{\mathcal{L}}$ . Then the augmented  $\mathcal{D}_{\mathcal{L}}$  with the labeled  $\mathcal{D}_{\mathcal{U}}$ , so-called  $\mathcal{D}_{\mathcal{L}+\mathcal{U}}$ , is used for the further iteration.
- (ii)  $T_{p_{0.5}}$ : The class probability of the  $\mathcal{D}_{\mathcal{U}}$  is equally assigned to  $\mathcal{D}_{\mathcal{L}}$  and used as an initial probability. In this approach, the  $\mathcal{D}_{\mathcal{L}+\mathcal{U}}$  can be derived earlier and integrated into training process for the first iteration. Therefore, in the next iteration, the  $\mathcal{D}_{\mathcal{U}}$  is not strictly guided by the labeled data. The revision process is probably the same manner to the previous method by combining the both dataset  $\mathcal{D}_{\mathcal{L}+\mathcal{U}}$  for the further iteration.
- (iii)  $T_{p_{random}}$ : Similarly, the initial probability of  $\mathcal{D}_{\mathcal{U}}$  is assigned randomly rather than the fixed value of 0.5. The degree of likelihood for each label can be varied from 0 to 1 whereas the total probability of ADR and IND labels equals to 1.

In order to evaluate our proposed method, three types of text representation across three parameters of unlabeled data incorporation are investigated. Finally, our proposed methods and its enhancement, MIL-dEM-S-S (supervised learning) and MIL-dEM-T-S

---

**Algorithm 3:** Pseudo-code for our proposed EM with BN dependent representation (dEM)

---

**Input:**  
 $|C|$  = the number of labels  
 $T$  = the maximum number of iteration  
 $\gamma_1, \gamma_2, \dots, \gamma_{|x_i|!}; \sum_{j=1}^{|x_i|!} \gamma_j = 1$   
 $\beta_1, \beta_2; \sum_{j=1}^2 \beta_j = 1$   
 $\alpha_1, \alpha_2, \dots, \alpha_4; \sum_{j=1}^4 \alpha_j = 1$   
**Output:**  $\theta$  parameter

```

1  $t \leftarrow 0$ 
2  $\theta = \{p^{(t+1)}(w_q, c_k), p^{(t+1)}(w_q, w_r, c_k), p^{(t+1)}(w_q, w_r, w_s, c_k), p^{(t)}(c_k)\}; \sum_{k=1}^{|C|} p^{(t)}(c_k) = 1$ 
3 repeat
4   for  $i = 1$  to  $n$  do
5     E-Step:
6       Estimate model probability :  $p^{(t)}(y_i = c_k | \mathbf{x}_i)$  (Eq.5.21)
7     M-Step:
8       Update class-conditional probability :  $p^{(t+1)}(w_q | c_k)$  (Eq.5.22)
9        $p^{(t+1)}(w_q | w_r, c_k)$  (Eq.5.23)
10       $p^{(t+1)}(w_q | w_r, w_s, c_k)$  (Eq.5.24)
11      Update class probability :  $p^{(t+1)}(c_k)$  (Eq.5.25)
12    $t \leftarrow t + 1$ 
13 until convergence or  $t = T$ 

```

---

methods (transductive learning), are compared to TSVM and three MIL models, MISVM, MINB, and MILR, which are implemented in WEKA [143].

## 5.4 Evaluation

In order to estimate the performance of the proposed method, the *hold-out evaluation* is conducted through the  $k$ -fold cross-validation whereas  $k = 5$ . The three main measures as defined by Eq.3.2-3.4 (see Section 3.3.2), i.e., *precision*, *recall* and *F1*, are used for model evaluation, while the positive class in our experiments is *ADR* label. The MetaMap Java API for named entity recognition, Stanford CoreNLP Java API for OpenIE, and implement Python program for EM-based methods. For model comparison, WEKA Java-based software and SVM<sup>light</sup><sup>4</sup>, which is implemented in C programming language, are executed on Mac OS with Intel Core i5 processor running at 2.5 GHz and 8 GB of physical memory. The experiment is conducted for five main experiments in order to evaluate the effectiveness of the proposed method; (i) the key phrasal patterns analysis, (ii) the evaluation on the effectiveness of the key phrasal patterns (iii) the effectiveness of the pattern-based feature with MIL-iEM and MIL-dEM, (iv) the evaluation on overall performance with advanced machine learning methods and (v) the evaluation on robustness of unlabeled data incorporation.

### 5.4.1 Data

The proposed framework is examined on the unstructured texts in EMR of Intensive Care Unit which is derived from MIMIC-III [39] (see Section 3.2.1). The data is freely available at *PhysioNet*<sup>5</sup> and is accessed on Apr 25, 2016 with the version 1.3. The over 58,000 hospital admissions for 38,645 adults and 7,875 neonates are presented in the data source with spanning up to 12 years from June 2001. A large scale of nearly 1.6 million sentences is extracted and used as the text corpus.

Consequently, the corpus is mapped to drug-event pairs that are reported in SIDER and DrugBank databases. All drug-event pairs that are appeared in both databases

---

<sup>4</sup><http://svmlight.joachims.org>

<sup>5</sup><https://mimic.physionet.org>

Table 5.3: The list of parameters for assessment

Parameter group	Parameter type	Parameter subtype	Parameter name	Variable name
Document representation	Feature representation	Syntactic lemmatization	Syntactically lemmatized lexicon	$L$
			Surface lexicon	$S$
		Pattern granularity	Phrase form	$P$
			Word form	$W$
	Pattern-weighting models	Bernoulli	Binary	$B$
			TF (term frequency)	$TF$
		Multinomial	TFIDF (TF-inverse document frequency)	$TFIDF$
Model assumption	Independent assumption	EM with Naïve Bayes		$iEM$
	Dependency representation assumption	EM with Bayesian Network		$dEM$
Model decision method	Soft decision making			$S$
	Hard decision making			$H$
Learning method	Supervised learning			$SL$
	Transductive learning	Initial weight method for unlabeled data	Supervised model	$T_{p_{ML}}$
			Equal probability	$T_{p_{0.5}}$
			Random probability	$T_{p_{random}}$

will be removed in order to avoid semantic ambiguity. Lastly, the 1,543 sentences corresponding drug-event pairs eventually is used as the training data.

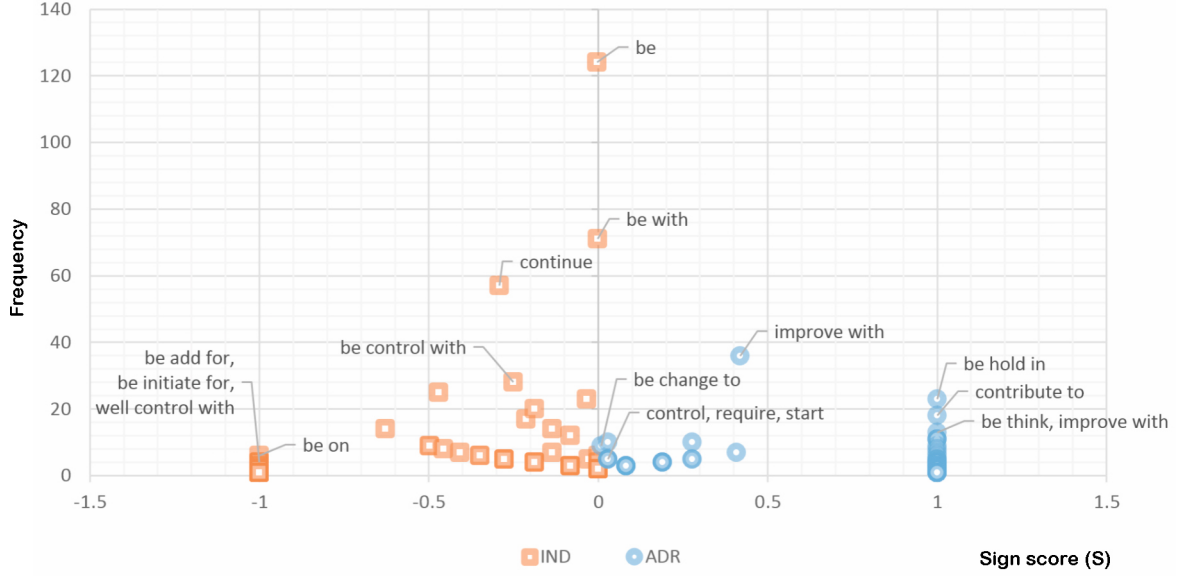


Fig. 5.4: The key phrasal pattern plot by adjusted entropy score vs. frequency.

### 5.4.2 Key phrasal patterns analysis

The discovered key phrasal patterns are analyzed to investigate the degree of characterization of relation labels. Given a key phrasal pattern, the pattern score ( $S$ ) is computed by performing the conditional entropy ( $H$ ) inversion and polarity adjustment as shown in Eq.5.26 to visualize the performance of the extracted key phrasal patterns.

$$\begin{aligned}
 H &= -p(ADR \mid pattern) \log_2 p(ADR \mid pattern) - p(IND \mid pattern) \log_2 p(IND \mid pattern) \\
 S &= SIGN(0.5 - p(IND \mid pattern)) \times (1 - H)
 \end{aligned}
 \tag{5.26}$$

From Figure. 5.4, the  $x$  axis exhibits the pattern score with polarity whereas the  $score > 0$  represents the distribution of pattern relevant ADR (blue circle marker), the  $score < 0$  represents the distribution of pattern relevant IND (orange square marker) and  $score = 0$  indicates no relevance between pattern and both labels. The  $y$  axis is the frequency of patterns that appear in the clinical texts. A pattern that is located far

from the middle line (score 0) and closed to the top left or the top right corners expresses the high effectiveness of semantic discrimination ability relevant relation labels. For example, the key phrasal patterns “*be-hold-in*”, “*contribute-to*”, “*be-think*”, “*improve-with*” are strongly relevant to ADR label and “*be-add-for*”, “*be-initial-for*”, “*be-on*” are rather associated to IND. Opposite to the key phrasal patterns “*be*” and “*be-with*” are presented near the middle line in the graph that indicates the fuzziest patterns.

Additionally, the figure clearly illustrates that the patterns relevant ADR are more efficient than the pattern relevant IND, the small number of ADR patterns are located nearby the original point and the most of ADR patterns are placed with spread distance. On the one hand, patterns relevant IND are presented to dense at the location which is nearly zero score and zero frequency. Table 5.4 presents the example of the sentences that are relevant to key phrasal patterns and pattern direction. Finally, the key phrasal patterns that have their pattern score over than threshold are selected for the further process.

### 5.4.3 Evaluation on the Effectiveness of the Key Phrasal Pattern-Based Feature

The comparison of the multiple feature types across varying of initial weighting of relation labels for unlabeled data incorporation throughout the MIL-iEM is assessed in order to examine the effectiveness of the pattern-based feature. The experiments are divided into two parts based on the decision methods in EM algorithm. The former refers to *soft decision making* (MIL-iEM-S) that the predicted result is directly yielded by the estimated class probability. The latter is so-called *hard decision making* (MIL-iEM-H) that the predicted outcome is considered on the cut off value of the probability and assigned class label instead of the likelihood ratio. The experimental setting is performed on the independent assumption through MIL-iEM model.

Table 5.5 expresses an assessment of nine text representation across three alternative document weighting model and three initial weighting method for unlabeled data  $\mathcal{D}_u$  based on *soft decision making*. In the table, the proposed phrasal pattern-based feature is expressed in the top 4 of each experimental setting, in other words, the combination

Table 5.4: An example of relevant sentences of drug-event (d, p, e) pairs.

Drugs (d)	Key Phrasal Patterns (p)	Events (e)	Pattern Direction	Sentences
<b>ADR</b>				
C0020261 (Hydrochlorothiazide)	be-hold-in	C0020625 (Hyponatremia)	d→e	However the patient’s sodium was 131 on discharge thus the Patient’s <b>HCTZ<sub>drug</sub> was-held-in<sub>pattern</sub></b> the setting of <b>hyponatremia<sub>event</sub></b> .
C0000970 (Acetaminophen)	be-think	C0002871 (Anemia)	e→d	Her <b>anemia<sub>event</sub> is-thought<sub>pattern</sub></b> to be due to direct effects of <b>acetaminophen<sub>drug</sub></b> on marrow or indirect via kidneys.
<b>IND</b>				
C0020223 (Hydrallazine)	be-give-for	C0020538 (Hypertension)	d→e	<b>Hydrallazine<sub>drug</sub> 20mg IV was-given-for<sub>pattern</sub></b> isolated episode of <b>hypertension<sub>event</sub></b> and emesis ensued.
C0043031 (Warfarin)	be-initiate-for	C0004238 (Atrial fibrillation)	e→d	<b>Warfarin<sub>drug</sub> was-initiated-for<sub>pattern</sub></b> his <b>atrial-fibrillation<sub>event</sub></b> with an initial heparin bridge.
<b>Pattern Direction:</b> d→e is drug-event, e→d is event-drug				

of text feature that starts with  $N$  such as  $N-S-P$  or  $N-S-W$  etc. As the result, the phrasal pattern-based feature has outperformed context inclusion feature (text feature that starts with  $C$ ). The highest F-score value, 0.841, is resulted by  $N-S-P$  feature using  $T_{p_{0.5}}$  with  $tf$  document model which has outperformed the baseline  $BOW$  up to 4.4%. Moreover, BD and MD( $tf$ ) document weighting are slightly better performance than MD( $tf-idf$ ) for all types of initial weighting of  $\mathcal{D}_U$ . The similar results are found on *hard decision making* approach as well. From Table 5.6, the phrasal pattern-based feature performs better performance than all context-based feature. The MD( $tf-idf$ ) document representation with  $N-L-W$  feature using  $T_{p_{0.5}}$  obtains the highest performance of F-score 0.807 and 3.3% improvement from the baseline. It is noticed that the *hard decision making* results in poor performance when compares to the soft version.

Another finding derived by this experimental results is that the independent text assumption is improper for relation classification problem. This is because, in general, a guideline from the existing  $\mathcal{D}_L$  to  $\mathcal{D}_U$  should improve the model performance. In the contrast by derived results,  $T_{p_{M_L}}$  yielded the poor performance, particularly, in  $N-S-P$  feature and lead to Type II error which is recognized as a significant issue, especially, in the medical domain.

The performance comparison across the number of features is exhibited in Fig. 5.5. The top graph represents to *soft decision* method, MIL-iEM-S, and the bottom graph is *hard decision* method, MIL-iEM-H. The number of features is ranged from 737 to 2,034 dimensions and the number of  $BOW$  feature is 1,853 dimensions. From the graph, even though the proposed pattern-based features with MIL-iEM- $T_{p_{0.5}}$  and MIL-iEM- $T_{p_{random}}$  provide slightly different F1-score from  $BOW$  feature, their number of dimensions is less than half of  $BOW$ , especially  $S-W$  and  $L-W$  features. Therefore, the proposed pattern-based feature is more efficient than  $BOW$  feature due to the small number of features but yield similar model performance.

Accordingly, the experimental results confidently support that the simplified sentence using relation tuple of a drug, a key phrasal pattern, and an event is potential feature transformation for relation classification task. Moreover, the ignoring of insignificant contexts can reduce redundancy of feature and avoid computational time issue that is frequently caused by the curse of dimensions.



Fig. 5.5: The number of features for each type of pattern weighting method across F1-score.

Table 5.5: The effectiveness comparison on 5-fold cross validation of text representation across three types of document weighting using MIL-iEM with *soft decision making* (MIL-iEM-S).

Models	BD			MD ( <i>tf</i> )			MD ( <i>tf-idf</i> )		
MIL-iEM-S	P	R	F1	P	R	F1	P	R	F1
$T_{p_{M_L}}$									
<i>N-S-P</i>	0.858	0.308	0.454	0.858	0.308	0.454	0.857	0.307	0.452
<i>N-S-W</i>	0.879	0.599	0.712	0.873	0.600	0.711	0.871	0.589	0.703
<i>N-L-P</i>	0.890	0.460	0.606	0.890	0.460	0.606	0.882	0.451	0.597
<i>N-L-W</i>	0.868	0.609	<b>0.716</b>	0.863	0.611	<b>0.716</b>	0.873	0.604	<b>0.714</b>
<i>C-S-P</i>	0.824	0.580	0.681	0.824	0.573	0.676	0.819	0.578	0.678
<i>C-S-W</i>	0.820	0.613	0.701	0.841	0.617	0.712	0.835	0.617	0.709
<i>C-L-P</i>	0.833	0.575	0.680	0.837	0.573	0.680	0.834	0.569	0.677
<i>C-L-W</i>	0.824	0.615	0.704	0.843	0.619	0.714	0.835	0.617	0.709
<i>BOW</i>	0.755	0.624	0.683	0.780	0.628	0.696	0.765	0.624	0.687
$T_{p_{0.5}}$									
<i>N-S-P</i>	0.845	0.836	<b>0.840</b>	0.844	0.838	<b>0.841</b>	0.846	0.830	<b>0.838</b>
<i>N-S-W</i>	0.783	0.792	0.788	0.784	0.801	0.792	0.787	0.743	0.764
<i>N-L-P</i>	0.840	0.816	0.828	0.840	0.816	0.828	0.836	0.799	0.817
<i>N-L-W</i>	0.785	0.797	0.791	0.796	0.799	0.798	0.777	0.714	0.744
<i>C-S-P</i>	0.719	0.931	0.811	0.774	0.896	0.831	0.766	0.896	0.826
<i>C-S-W</i>	0.721	0.938	0.815	0.784	0.874	0.827	0.774	0.889	0.828
<i>C-L-P</i>	0.726	0.936	0.818	0.782	0.891	0.833	0.765	0.901	0.827
<i>C-L-W</i>	0.724	0.927	0.813	0.790	0.874	0.830	0.784	0.874	0.827
<i>BOW</i>	0.692	0.927	0.793	0.749	0.850	0.797	0.735	0.861	0.793

*Continued on next page*

Table 5.5 – Continued from previous page

Models	BD			MD ( <i>tf</i> )			MD ( <i>tf-idf</i> )		
MIL-iEM-S	P	R	F1	P	R	F1	P	R	F1
$T_{Prandom}$									
$N-S-P$	0.840	0.836	<b><u>0.838</u></b>	0.842	0.834	<b><u>0.838</u></b>	0.841	0.819	<b><u>0.830</u></b>
$N-S-W$	0.773	0.790	0.782	0.782	0.792	0.787	0.782	0.732	0.756
$N-L-P$	0.833	0.830	0.832	0.833	0.828	0.831	0.835	0.801	0.818
$N-L-W$	0.783	0.796	0.789	0.799	0.805	0.802	0.778	0.710	0.742
$C-S-P$	0.729	0.922	0.814	0.766	0.883	0.820	0.765	0.887	0.822
$C-S-W$	0.724	0.927	0.813	0.787	0.863	0.823	0.778	0.874	0.823
$C-L-P$	0.731	0.931	0.819	0.775	0.876	0.823	0.764	0.885	0.820
$C-L-W$	0.724	0.925	0.813	0.786	0.867	0.825	0.784	0.863	0.822
$BOW$	0.691	0.900	0.782	0.748	0.834	0.789	0.734	0.841	0.784

B: binary frequency, TF: term frequency, TFIDF: term frequency-inverse document frequency.

Table 5.6: The effectiveness comparison on 5-fold cross validation of text representation across three types of document weighting using MIL-iEM with *hard decision making* (MIL-iEM-H).

Models	BD			MD ( <i>tf</i> )			MD ( <i>tf-idf</i> )		
MIL-iEM-H	P	R	F1	P	R	F1	P	R	F1
$T_{p_{ML}}$									
$N-S-P$	0.856	0.327	0.473	0.856	0.327	0.473	0.858	0.330	0.477
$N-S-W$	0.871	0.651	0.745	0.863	0.642	0.736	0.873	0.650	0.745
$N-L-P$	0.887	0.500	0.639	0.887	0.500	0.639	0.881	0.498	0.636
$N-L-W$	0.866	0.659	<b>0.748</b>	0.863	0.653	<b>0.744</b>	0.868	0.662	<b>0.752</b>
$C-S-P$	0.805	0.588	0.679	0.809	0.588	0.681	0.799	0.586	0.676
$C-S-W$	0.798	0.626	0.701	0.812	0.624	0.706	0.807	0.624	0.704
$C-L-P$	0.810	0.582	0.677	0.820	0.582	0.681	0.814	0.582	0.679
$C-L-W$	0.806	0.628	0.706	0.817	0.626	0.709	0.813	0.626	0.707
$BOW$	0.744	0.642	0.690	0.730	0.646	0.685	0.726	0.642	0.682
$T_{p_{0.5}}$									
$N-S-P$	0.620	0.985	0.761	0.620	0.985	0.761	0.621	0.985	0.762
$N-S-W$	0.686	0.971	0.804	0.683	0.969	0.802	0.691	0.967	0.806
$N-L-P$	0.651	0.974	0.781	0.652	0.974	0.781	0.651	0.973	0.780
$N-L-W$	0.693	0.962	<b>0.805</b>	0.689	0.960	<b>0.802</b>	0.696	0.960	<b>0.807</b>
$C-S-P$	0.630	0.984	0.768	0.641	0.973	0.772	0.643	0.974	0.775
$C-S-W$	0.634	0.984	0.771	0.645	0.969	0.775	0.649	0.971	0.778
$C-L-P$	0.632	0.982	0.769	0.639	0.971	0.771	0.646	0.973	0.776
$C-L-W$	0.639	0.987	0.776	0.645	0.971	0.775	0.652	0.973	0.780
$BOW$	0.632	0.973	0.766	0.646	0.962	0.773	0.649	0.960	0.774

*Continued on next page*

Table 5.6 – *Continued from previous page*

Models	BD			MD ( <i>tf</i> )			MD ( <i>tf-idf</i> )		
MIL-iEM-H	P	R	F1	P	R	F1	P	R	F1
$T_{p_{random}}$									
$N-L-P$	0.645	0.755	0.696	0.644	0.754	0.695	0.630	0.757	0.688
$N-L-W$	0.666	0.825	<b><u>0.737</u></b>	0.649	0.834	0.730	0.640	0.856	0.732
$N-S-P$	0.668	0.814	0.734	0.668	0.814	<b><u>0.734</u></b>	0.656	0.841	<b><u>0.737</u></b>
$N-S-W$	0.657	0.827	0.732	0.642	0.823	0.722	0.641	0.863	0.736
$C-L-P$	0.686	0.746	0.715	0.674	0.790	0.728	0.683	0.779	0.728
$C-L-W$	0.681	0.757	0.717	0.666	0.790	0.723	0.680	0.790	0.731
$C-S-P$	0.692	0.766	0.727	0.671	0.794	0.727	0.687	0.790	0.735
$C-S-W$	0.687	0.755	0.719	0.666	0.779	0.718	0.676	0.785	0.726
$BOW$	0.655	0.746	0.698	0.666	0.801	0.727	0.675	0.794	0.730

B: binary frequency, TF: term frequency, TFIDF: term frequency-inverse document frequency.

#### 5.4.4 Evaluation on the Effectiveness of MIL-dEM-SL and MIL-dEM-T

In this experiment, the comparison between the proposed method based on supervised learning (MIL-dEM-SL) and transductive learning (MIL-dEM-T) across varying parameters such as feature types, pattern-weighting models, and the initial weight methods for unlabeled data incorporation are examined. The proposed method is based on dependency representation of texts, and the posterior estimation is based on the interpolation of Markov property. The experiment is set up with the supervised learning-based model, and three transductive learning-based models with different initial weight methods of  $\mathcal{D}_{\mathcal{U}}$  incorporation. The two types of pattern-based features such as surface lexicon-based ( $N-S-P$ ) and syntactically lemmatized lexicon-based ( $N-L-P$ ) are used for examination. The parameter tuning is also performed for all approaches.

As the results in Table 5.7, among transductive learning models, the performance of  $N-S-P$  feature is slightly different from  $N-L-P$  feature for all models. The simply binary ( $B$ ) weighting model presents the higher F1-score over  $TF$  and  $TFIDF$ . Moreover, MIL-dEM-S- $T_{p_{ML}}$  model exhibits the higher performance than the fuzzy guideline by MIL-dEM-S- $T_{p_{0.5}}$  and MIL-dEM-S- $T_{p_{random}}$  models for all evaluation matrices.

On the other hand, the F1-score of MIL-dEM-SP-S-SL surface lexicon-based feature is better than MIL-dEM-LP-S-SL syntactically lemmatized lexicon-based feature with 1% and 0.8% for  $TF$  and  $TFIDF$  weighting model respectively. Similarly, the F1-score of the pattern-based feature  $N-S-P$  across the three types of pattern-weighting model, i.e.,  $B$ ,  $TF$ ,  $TFIDF$ , models is also slightly different; 0.928 for MIL-dEM-SP-B-S-SL, 0.946 for MIL-dEM-SP-TF-S-SL, and 0.938 for MIL-dEM-SP-TFIDF-S-SL. Among models within MIL-dEM-S-SL setting, the highest F1-score is presented by  $TF$  weighting model with 0.946.

One of the interesting results shows that the unlabeled data incorporation is exhibited to increase the model performance. The highest effectiveness, 0.954 of F1-score, is presented by MIL-dEM-SP-B-S- $T_{p_{ML}}$  model which is the simply binary weighting model, and the model is shown 2.6% improvement over MIL-dEM-SP-B-S-SL, 1.6% improvement over MIL-dEM-SP-TFIDF-S-SL and 0.8% improvement over MIL-dEM-

SP-TF-S-SL, which is the best performance of our proposed supervised learning.

According to the result from the parameters optimization of our proposed method, the model performance is strongly relevant to the dependency representation of random variables as the following; (i) an event and the clinical outcome, (ii) a pattern, a drug and the clinical outcome. In the contrast, the model is shown the less relevance between a drug and an event, or a pattern and an event.

#### 5.4.5 Evaluation on Overall Performance with Advanced Machine Learning Methods

The comparison of our proposed method and advanced machine learning methods is presented in Table 5.8. The best models of each set of models are used for assessment. The well-known MIL methods i.e., MISVM, MINB, MILR are executed using WEKA. On the one hand, we customize the original TSVM using the source code from the author and incorporate the MIL assumption as discussed in the previous section (See section 2.5). We divide the discussion into three parts; the effectiveness of supervised learning model, the effectiveness of transductive learning model, and the overall performance.

Firstly, the experimental results among baseline supervised learning methods, i.e., MISVM-TFIDF, MINB-B, MILR-B, show that *BOW* feature works well for all MIL methods, conversely, the pattern-based feature *N-S-P* contributes a dramatic improvement when it is combined with our proposed method MIL-dEM-TF-S-SL. The *TFIDF* weighting model yields the high performance for MISVM with F1-score 0.901, while binary weighting model (*B*) is exhibited to improve the performance for MINB and MILR with F1-score 0.880 and 0.861 respectively. However, our proposed MIL-dEM-TF-S-SL with *N-S-P* feature outperforms all MIL methods and 4.5% F1-score better than the highest performance of advanced machine learning method which is resulted by MISVM-TFIDF with *BOW* feature. Although the precision of MIL-dEM-TF-S-SL with *N-S-P* feature is slightly lower than MISVM-TFIDF with *BOW* but the recall is significantly improved. Accordingly, our proposed method contributes to reducing the Type II error which is always considered in the medical domain.

Table 5.7: The effectiveness of MIL-dEM-S-SL and MIL-dEM-S-T comparison across three types of initial weight on 5-fold cross validation with *soft decision making*.

Models		$B$			$TF$			$TFIDF$		
		P	R	F1	P	R	F1	P	R	F1
<b>MIL-dEM-S-SL<sup>1</sup></b>										
<b>Supervised Learning</b>	$N-S-P$	0.883	<u>0.978</u>	<u>0.928</u>	<u>0.904</u>	<u>0.993</u>	<u>0.946</u>	<u>0.890</u>	<u>0.993</u>	<u>0.938</u>
	$N-L-P$	<u>0.896</u>	0.962	<u>0.928</u>	0.898	0.978	0.936	0.889	0.976	0.930
<b>MIL-dEM-S-T<sup>2</sup><sub><math>p_{ML}</math></sub></b>										
	$N-S-P$	<u>0.934</u>	<u>0.975</u>	<u>0.954</u>	0.901	<u>0.942</u>	<u>0.921</u>	<u>0.881</u>	<u>0.951</u>	<u>0.915</u>
	$N-L-P$	0.926	0.962	0.944	<u>0.919</u>	0.916	0.918	0.875	0.945	0.909
<b>MIL-dEM-S-T<sup>3</sup><sub><math>p_{0.5}</math></sub></b>										
<b>Transductive Learning</b>	$N-S-P$	0.839	<u>0.907</u>	<u>0.872</u>	0.635	<u>0.925</u>	0.754	0.686	<u>0.916</u>	0.784
	$N-L-P$	<u>0.850</u>	0.889	0.869	<u>0.663</u>	0.900	<u>0.763</u>	<u>0.714</u>	0.887	<u>0.791</u>
<b>MIL-dEM-S-T<sup>4</sup><sub><math>p_{random}</math></sub></b>										
	$N-S-P$	0.830	<u>0.889</u>	<u>0.859</u>	0.581	0.607	0.594	0.647	<u>0.682</u>	0.664
	$N-L-P$	<u>0.843</u>	0.865	0.854	<u>0.597</u>	<u>0.619</u>	<u>0.608</u>	<u>0.657</u>	0.679	<u>0.668</u>

<sup>1,2</sup> $\gamma = [0.45 \ 0.02 \ 0.45 \ 0.02 \ 0.04 \ 0.02], \beta = [0.97 \ 0.02 \ 0.01 \ 0.00], \alpha = [0.10 \ 0.90]$

<sup>3,4</sup> $\gamma = [0.45 \ 0.02 \ 0.45 \ 0.02 \ 0.04 \ 0.02], \beta = [1.00 \ 0.00 \ 0.00 \ 0.00], \alpha = [0.50 \ 0.50]$

B: binary frequency, TF: term frequency, TFIDF: term frequency-inverse document frequency.

Secondly, the comparison among transductive learning methods, the *BOW* feature with TSVM-B is shown to achieve F1-score of 0.889, while applying of the pattern-based feature  $N-S-P$ , its performance is presented to degrade around 2%. Conversely, the pattern-based feature  $N-S-P$  with MIL generative method is exhibits to enhance the effectiveness of the models. The accuracy of MIL-iEM-TF-S- $T_{p_{0.5}}$  model increases up to 6.3% when deploys the pattern-based feature instead of *BOW* feature.

Lastly, the overall evaluation, the generative models with dependency representation, i.e., MIL-dEM-TF-S-SL and MIL-dEM-B-S- $T_{p_{ML}}$ , outperform for all models. The highest performance is exhibited by our transductive learning MIL-dEM-B-S- $T_{p_{ML}}$  method with 0.934 precision, 0.975 recall, 0.954 F1-score and 0.949 accuracy respectively. Moreover, improving the generative model by substitute assumption of word-dependency MIL-dEM-B-S- $T_{p_{ML}}$  model to word-indenpendency MIL-iEM-TF-S- $T_{p_{0.5}}$  model is shown to dramatically improve 11.3% F1-score and 12.2% accuracy.

From multiple aspects assessment, the experimental results confidently support that our proposed methods, MIL with the two generative models, is comparative advantage in relation classification with the high performance. The proposed pattern-based feature contributes to reduce the curse of dimension issue and preserve texts dependency structure. The incorporation of a generative model with proper model assumption and transductive learning can potentially estimate the distribution of patterns relevant harmful or beneficial event of drug usage with the high precision and recall. Our proposed method can provide the supporting evidence based on the relevant clinical sentence rather than only prediction of result which is expected to further assist a professional medical for decision making on treatment or diagnosis process.

#### 5.4.6 Evaluation on Effect of Unlabeled Data Incorporation

The durable of the proposed method is examined by varying the ratio of unlabeled and labeled data. Two document representation,  $N-S-P$  and  $N-L-P$ , with three types of document weighting model, *BD*, *TF*, *TFIDF* are evaluated. Such parameters are constructed then learned on MIL-dEM-S- $T_{p_{ML}}$  model. The numbers of unlabeled data are varied from 4000, 10000 and 50000 examples. The F1-measure is used for

Table 5.8: The comparison of overall performance among MIL-dEM-SL, MIL-dEM-T, advanced machine learning methods, and MIL-iEM-T using 5-fold cross validation.

Models	BOW			N-S-P				
	P	R	F1	Acc.	P	R	F1	Acc.
Supervised Learning								
MIL-dEM-TF-S-SL <sup>1</sup>	-	-	-	-	<u>0.904</u>	<u>0.993</u>	<u>0.946</u>	<u>0.939</u>
MISVM-TFIDF <sup>2</sup>	<u>0.918</u>	0.885	<u>0.901</u>	<u>0.895</u>	0.799	0.733	0.765	0.735
MINB-B	0.864	<u>0.896</u>	0.880	0.867	0.619	0.701	0.744	0.691
MILR-B <sup>3</sup>	0.869	0.852	0.861	0.850	0.718	0.783	0.749	0.692
Transductive Learning								
MIL-dEM-B-S-T <sup>4</sup> <sub>p<sub>ML</sub></sub>	-	-	-	-	<u>0.934</u>	<u>0.975</u>	<u>0.954</u>	<u>0.949</u>
TSVM-B	<u>0.898</u>	<u>0.881</u>	<u>0.889</u>	0.881	0.873	0.865	0.869	0.859
MIL-iEM-TF-S-T <sub>p<sub>0.5</sub></sub>	0.749	0.850	0.797	0.764	0.844	0.838	0.841	0.827

<sup>1,4</sup> $\gamma = [0.45 \ 0.02 \ 0.45 \ 0.02 \ 0.04 \ 0.02], \beta = [0.97 \ 0.02 \ 0.01 \ 0.00], \alpha = [0.10 \ 0.90]$

<sup>2</sup> *polynomial kernel,  $C = 10$*

<sup>3</sup> *collective MI assumption, geometric mean for posteriors*

B: binary frequency, TF: term frequency, TFIDF: term frequency-inverse document frequency.

comparison. Figure 5.6 exhibits the performance of the proposed method, dEM, across numbers of unlabeled data. The experimental results show that the trends of F1 score are slightly reduced when numbers of unlabeled data is increased. The  $N-S-P$  with binomial distribution provides the higher performance than others representation. The increasing of unlabeled data around 2.5 times, F1 score reduces around 0.01, and the increasing of unlabeled data around 12.5 times, F1 score reduces around 0.02. On the other hand, the model with  $N-L-P$  with TF is more robustness. The increasing number of unlabeled data 12.5 times, the trend of F1 score is slightly changed around 0.005.

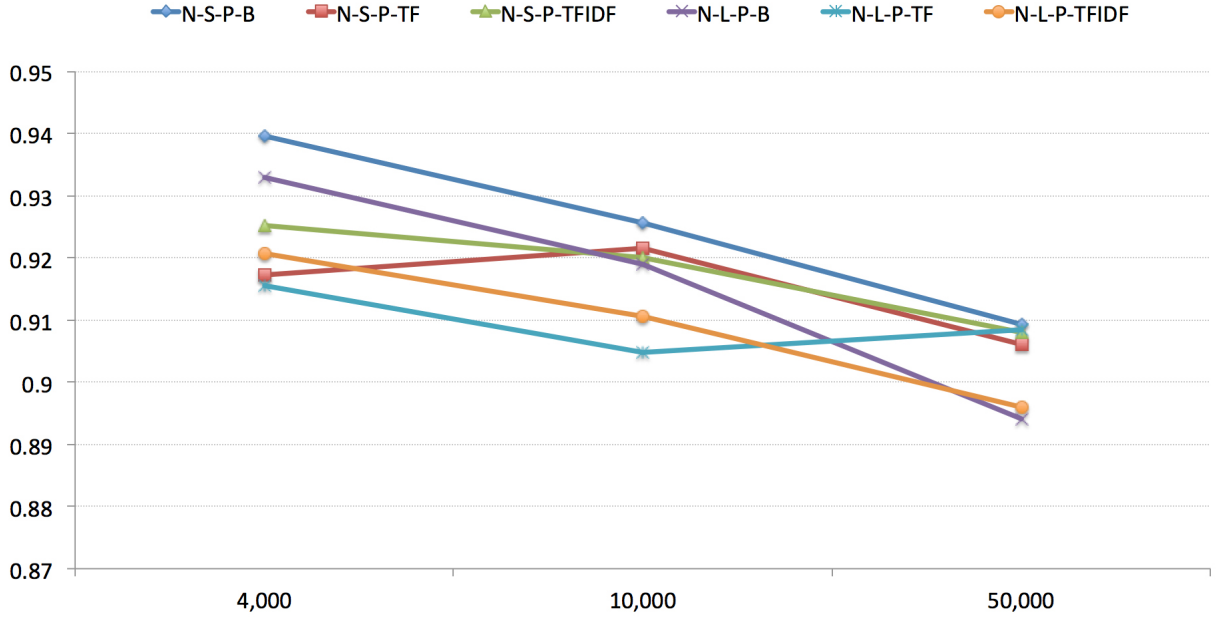


Fig. 5.6: F1-score vs. numbers of unlabeled data

## 5.5 Summary

This chapter presents a framework of distant supervision with multiple instance learning and transductive inference for detecting hidden adverse reaction in clinical texts. The work aims to deal with two main difficulties; (i) the limitation of hand-labeled data, and (ii) intractable processing of large-scale unstructured clinical texts.

The first issue is coped with distant supervision paradigm by knowledge bases incor-

poration. Therefore, either ADR or IND relation label can be automatically assigned to each drug-event pair and use as labeled examples. For the second issue, the key phrasal pattern-based feature is investigated to present semantic comprehension of a sentence and proposed alternative parameters learning of a generative model using dependency representation model assumption. However, such training data set derived by distant supervision is formed as the instance-level, while the predictive goal is focused on the entity-level. Therefore, MIL paradigm is involved into the framework. The collected statistics from the tagged drug-event pairs are used to examine the semantic distribution relevant to ADR and IND. Exploiting EM algorithm as the base model for our supervised learning and transductive learning, it is helpful to estimate the probability of an unknown relation of given drug-event pair and then classify this relation to either ADR or IND. From the experimental results on multiple assessments, we found three significant findings.

Firstly, the pattern-based feature contributes to improving model performance of generative models. The MIL-iEM-SP-TF-S- $T_{p_{0.5}}$  model is shown to achieve the highest performance among all MIL-iEM-based methods with 0.844 precision, 0.838 recall and 0.841 F1-score, and the model provides the outstanding improvement over the traditional *BOW* method, MIL-iEM-BOW-TF-S- $T_{p_{0.5}}$  model, up to 4.4% F1-score.

The second potential result, the traditional assumption of word-independency is rather improper for natural clinical texts. Therefore, we tackle such fundamental problem by integrating Markov assumption on dependency representation of texts in order to estimate the prior probability and likelihood probability in a generative model. Given the same set of the pattern-based input features, the performance of MIL-dEM model is dramatically improved from MIL-iEM model. The MIL-dEM-SP-B-S- $T_{p_{ML}}$  model exhibits the improvement over MIL-iEM-SP-B-S- $T_{p_{0.5}}$  up to 8.9% precision, 13.9% recall and 11.4% F1-score.

Lastly, the incorporation of unlabeled data  $\mathcal{D}_U$  and labeled one  $\mathcal{D}_L$  using MIL-dEM-SP-B-S- $T_{p_{ML}}$  model achieves the highest efficiency with 0.954 F1-score. In addition, our proposed MIL-dEM-SP-B-S- $T_{p_{ML}}$  model also outperforms the advanced machine learning methods by F1-score improvement up to 5.3% of MISVM-BOW-TFIDF, 7.4% of MINB-BOW-B, 9.3% of MILR-BOW-B, 6.5% of TSVM-BOW-B and 11.3% of MIL-

iEM-SP-TF-S-T<sub>p<sub>0.5</sub></sub>.

However, our work presents some limitations that can contribute to support further improvement of the framework. The projection from distant supervision to corpus currently is employed by MetaMap tools and can be improved by an advanced method such as word embedding to increase high potential entity-level relation for instance examples. The key phrasal patterns extraction in the current work is scoped by the sentence boundary, but a drug and an event possibly associate throughout across different sentences. This issue would be challenged by the co-reference problem. Even though the discovered key phrasal patterns provides the significant role for relation classification but the number of patterns is rather limited and probably encounters the problem of out of vocabulary (OOV) when applies the framework with a huge unseen data. Therefore, the semantic representation is the promising method to increase the number of key phrasal patterns.

# Chapter 6

## Conclusion

This dissertation studies on Text Mining for information extraction, more specifically, relation extraction. The relation extraction, firstly, aims to find a candidate of entity pairs that is possibly formed a relation, then the process of classification such relation is probably employed. The former task is so-called relation detection, and the latter one is namely relation classification. In other words, the relation detection targets to identify a link (or connection) between any entity pair, and relation classification aims to provide more information of relation type (or relation label). Such semantic relation is very useful for comprehension, especially, in the medical domain.

To deal with relation extraction, there are multidisciplinary such as supervised learning, unsupervised learning, and semi-supervised learning. While supervised learning achieves the high performance for classification, the drawback is relevant to the insufficient for data training (such as a rare case for training data) or encounter a large volume of unlabeled data for instance labeling. Particular, data labeling is recognized as difficult, domain-dependent, expensive and time-consuming.

Therefore, this thesis addresses two fundamental problems; (i) The lack of domain experts for labeling examples, especially, in a large volume of unlabeled data; (ii) The intractable processing of unstructured text, particularly, clinical text. To this end, firstly, the thesis presents the generic framework as a solution for semantic relation extraction from text. Moreover, the framework can also be applied in any domain with certain assumption. Secondly, the thesis introduces the efficient parameters estimation

in a generative model that argues the traditional text assumption. This contribution can help to dramatically improve the performance of the model. Lastly, the thesis contributes to examine the multiple approaches of unlabeled data augmentation in order to deal with a large-scale of data with the effectiveness.

## Future works

In this thesis, there are many rooms for further improvement as the following topics:

- **Relation detection:** the considering only the named entity of a drug and an event entities that are found in the same sentence is rather strict. The co-reference extraction could be promising for making relax assumption. A drug or an event might form a relation, even though they are found in the different sentence. Therefore an indirect relation can be considered.
- **Out of vocabulary (OOV):** the extracted phrase pattern is used to bootstrapping the new pattern by means of pattern matching method which is limited to the number of discovered phrase pattern. It is also known as *out of vocabulary* issue. This can be improved by novel feature representation such as word embedding by considering of the generalization for pattern matching. This is expected to improve retrieval rate.
- **The complexity of parameter tuning:** the proposed method MIL-dEM seems to fall into infeasible for parameter tuning due to the number of parameters. This issue can be improved through the coefficient learning which is dynamic weighting based on the data distribution for each iteration.
- **Extensive knowledge base:** regarding the distant supervision, the quality of model depends on the source of a knowledge base for projection. In this thesis, the current problem is based on a binary relation classification. However, in the real world, the pattern between two entities might represent more than two semantic labels. It is highly recommended to integrate new sources of data in order to improve discriminative patterns.

# Bibliography

- [1] I. H. Witten, “Adaptive text mining: inferring structure from sequences,” *Journal of Discrete Algorithms*, vol. 2, no. 2, pp. 137–159, 2004.
- [2] J. Piskorski and R. Yangarber, “Information Extraction: Past, Present and Future,” in *Multi-source, Multilingual Information Extraction and Summarization*, pp. 23–49, Springer, 2013.
- [3] A. Kothari, D. Rudman, M. Dobbins, M. Rouse, S. Sibbald, and N. Edwards, “The use of tacit and explicit knowledge in public health: a qualitative study,” *Implementation Science*, vol. 7, no. 1, p. 1, 2012.
- [4] J. Lee, D. M. Maslove, and J. A. Dubin, “Personalized mortality prediction driven by electronic medical data and a patient similarity metric,” *PloS one*, vol. 10, no. 5, p. e0127428, 2015.
- [5] T. Tran, W. Luo, D. Phung, R. Harvey, M. Berk, R. L. Kennedy, and S. Venkatesh, “Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments,” *BMC psychiatry*, vol. 14, no. 1, p. 1, 2014.
- [6] E. H. Kennedy, W. L. Wiitala, R. A. Hayward, and J. B. Sussman, “Improved cardiovascular risk prediction using nonparametric regression and electronic health record data,” *Medical care*, vol. 51, no. 3, p. 251, 2013.
- [7] O. Frunza, D. Inkpen, and T. Tran, “A machine learning approach for identifying disease-treatment relations in short texts,” *IEEE transactions on knowledge and data engineering*, vol. 23, no. 6, pp. 801–814, 2011.

- [8] L. Tari, S. Anwar, S. Liang, J. Cai, and C. Baral, “Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism,” *Bioinformatics*, vol. 26, no. 18, pp. i547–i553, 2010.
- [9] I. Segura-Bedmar, P. Martinez, and C. de Pablo-Sánchez, “Using a shallow linguistic kernel for drug–drug interaction extraction,” *Journal of biomedical informatics*, vol. 44, no. 5, pp. 789–804, 2011.
- [10] WHO, “International drug monitoring: the role of national centres,” Tech. Rep. 498, Tech Rep Ser WHO, 1972.
- [11] T.-B. Ho, L. Le, D. T. Thai, and S. Taewijit, “Data-driven approach to detect and predict adverse drug reactions,” *Current pharmaceutical design*, vol. 22, no. 23, pp. 3498–3526, 2016.
- [12] C. Friedman, “Discovering novel adverse drug events using natural language processing and mining of the electronic health record,” in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 1–5, Springer, 2009.
- [13] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. The MIT Press, 1st ed., 2010.
- [14] M. Craven, J. Kumlien, *et al.*, “Constructing biological knowledge bases by extracting information from text sources,” in *ISMB*, vol. 1999, pp. 77–86, 1999.
- [15] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL ’09, (Stroudsburg, PA, USA), pp. 1003–1011, Association for Computational Linguistics, 2009.
- [16] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL ’95, (Stroudsburg, PA, USA), pp. 189–196, Association for Computational Linguistics, 1995.

- [17] X. Zhu, Z. Ghahramani, J. Lafferty, *et al.*, “Semi-supervised learning using gaussian fields and harmonic functions,” in *ICML*, vol. 3, pp. 912–919, 2003.
- [18] G. Erkan, A. Özgür, and D. R. Radev, “Semi-supervised classification for extracting protein interaction sentences using dependency parsing.,” in *EMNLP-CoNLL*, vol. 7, pp. 228–237, 2007.
- [19] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
- [20] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.
- [21] T. Joachims, “Transductive inference for text classification using support vector machines,” in *ICML*, vol. 99, pp. 200–209, 1999.
- [22] M. Belkin and P. Niyogi, “Semi-supervised learning on riemannian manifolds,” *Machine learning*, vol. 56, no. 1-3, pp. 209–239, 2004.
- [23] I. Triguero, S. García, and F. Herrera, “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study,” *Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, 2015.
- [24] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, “Self-trained lmt for semisupervised learning,” *Computational intelligence and neuroscience*, vol. 2016, p. 10, 2016.
- [25] L. Didaci, G. Fumera, and F. Roli, “Analysis of co-training algorithm with very small training sets,” in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 719–726, Springer, 2012.
- [26] R. Feldman and I. Dagan, “Knowledge Discovery in Textual Databases (KDT).,” in *KDD*, vol. 95, pp. 112–117, 1995.

- [27] A.-H. Tan *et al.*, “Text mining: The state of the art and the challenges,” in *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, vol. 8, pp. 65–70, 1999.
- [28] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge university press, 2007.
- [29] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, and M. Tyson, “FASTU: A Finite-state Processor for Information Extraction from Real-world Text,” in *IJCAI*, vol. 93, pp. 1172–1178, 1993.
- [30] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, M. Kameyama, D. Martin, K. Myers, and M. Tyson, “SRI International FASTUS system: MUC-6 test results and analysis,” in *Proceedings of the 6th conference on Message understanding*, pp. 237–248, Association for Computational Linguistics, 1995.
- [31] R. Grishman, “The NYU System for MUC-6 or Where’s the Syntax?,” in *Proceedings of the 6th conference on Message understanding*, pp. 167–175, Association for Computational Linguistics, 1995.
- [32] F. Rinaldi, S. Clematide, H. Marques, T. Ellendorff, M. Romacker, and R. Rodriguez-Esteban, “OntoGene web services for biomedical text mining,” *BMC bioinformatics*, vol. 15, no. 14, p. S6, 2014.
- [33] R. Srihari, C. Niu, and W. Li, “A Hybrid Approach for Named Entity and Sub-Type Tagging,” in *Proceedings of the sixth conference on Applied natural language processing*, pp. 247–254, Association for Computational Linguistics, 2000.
- [34] F. Jenhani, M. S. Gouider, and L. B. Said, “A hybrid approach for drug abuse events extraction from twitter,” *Procedia Computer Science*, vol. 96, pp. 1032–1040, 2016.
- [35] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 1 ed., 2000. neue Auflage kommt im Frhjahr 2008.

- [36] R. Grishman, “Information extraction: Techniques and challenges,” in *Information extraction a multidisciplinary approach to an emerging information technology*, pp. 10–27, Springer, 1997.
- [37] M. A. Hearst, “Untangling Text Data Mining,” in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3–10, Association for Computational Linguistics, 1999.
- [38] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining,” in *Ldv Forum*, vol. 20, pp. 19–62, 2005.
- [39] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, 2016.
- [40] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris, “Text and Data Mining Techniques in Adverse Drug Reaction Detection,” *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 56, 2015.
- [41] E. J. Hovenga, *Health Informatics: An Overview*, vol. 151. Ios Press, 2010.
- [42] M. Liu, E. R. M. Hinz, M. E. Matheny, J. C. Denny, J. S. Schildcrout, R. A. Miller, and H. Xu, “Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records,” *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 420–426, 2013.
- [43] T. Penney, “Dictate a discharge summary,” *BMJ: British Medical Journal*, vol. 298, no. 6680, p. 1084, 1989.
- [44] S. Doan, N. Collier, H. Xu, P. H. Duy, and T. M. Phuong, “Recognition of medication information from discharge summaries using ensembles of classifiers,” *BMC medical informatics and decision making*, vol. 12, no. 1, p. 36, 2012.
- [45] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, “Automated Acquisition of DiseaseDrug Knowledge from Biomedical and Clinical Documents:

- An Initial Study,” *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87–98, 2008.
- [46] X. Wang, G. Hripcsak, and C. Friedman, “Characterizing environmental and phenotypic associations using information theory and electronic health records,” *BMC bioinformatics*, vol. 10, no. Suppl 9, p. S13, 2009.
  - [47] X. Wang, G. Hripcsak, M. Markatou, and C. Friedman, “Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study,” *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 328–337, 2009.
  - [48] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman, “Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions,” *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 413–419, 2013.
  - [49] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, “A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.
  - [50] Y. Li, P. B. Ryan, Y. Wei, and C. Friedman, “A Method to Combine Signals from Spontaneous Reporting Systems and Observational Healthcare Data to Detect Adverse Drug Reactions,” *Drug safety*, vol. 38, no. 10, pp. 895–908, 2015.
  - [51] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, “Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis,” *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.
  - [52] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, *et al.*, “Overview of the ShARe/CLEF eHealth Evaluation Lab 2013,” in *International Conference*

- of the Cross-Language Evaluation Forum for European Languages, pp. 212–231, Springer, 2013.
- [53] L. Duan, M. Khoshneshin, W. N. Street, and M. Liu, “Adverse drug effect detection,” *IEEE journal of biomedical and health informatics*, vol. 17, no. 2, pp. 305–311, 2013.
  - [54] M. Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X.-w. Chen, M. E. Matheny, and H. Xu, “Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs,” *Journal of the American Medical Informatics Association*, vol. 19, no. e1, pp. e28–e35, 2012.
  - [55] P. LePendou, Y. Liu, S. Iyer, M. R. Udell, and N. H. Shah, “Analyzing patterns of drug use in clinical notes for patient safety,” *AMIA Summits Transl Sci Proc*, vol. 2012, pp. 63–70, 2012.
  - [56] J. Zhao, A. Henriksson, and H. Boström, “Detecting Adverse Drug Events Using Concept Hierarchies of Clinical Codes,” in *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*, pp. 285–293, IEEE, 2014.
  - [57] A. Henriksson, M. Kvist, M. Hassel, and H. Dalianis, “Exploration of Adverse Drug Reactions in Semantic Vector Space Models of Clinical Text,” in *Proceedings of ICML Workshop on Machine Learning for Clinical Data Analysis*, 2012.
  - [58] H. Dalianis, M. Hassel, A. Henriksson, and M. Skeppstedt, “Stockholm epr corpus: A clinical database used to improve health care,” in *Swedish Language Technology Conference*, pp. 17–18, 2012.
  - [59] D. Yoon, M. Park, N. Choi, B. Park, J. Kim, and R. Park, “Detection of Adverse Drug Reaction Signals Using an Electronic Health Records Database: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) algorithm,” *Clinical pharmacology and therapeutics*, vol. 91, no. 3, p. 467, 2012.
  - [60] I. R. Edwards, “Spontaneous reporting of what? clinical concerns about drugs,” *British journal of clinical pharmacology*, vol. 48, no. 2, pp. 138–141, 1999.

- [61] M. Stephens, J. Talbot, and P. Waller, *Stephens' detection of new adverse drug reactions*. John Wiley & Sons, 2004.
- [62] R. D. Mann and E. B. Andrews, *Pharmacovigilance*. John Wiley & Sons, 2007.
- [63] R. M. Twyman, *Principles of proteomics*. Garland Science, 2013.
- [64] G. Barchet, "A brief overview of metabolomics: What it means, how it is measured, and its utilization," *The Science Creative Quarterly*, vol. 8, 2013.
- [65] R. L. Blaylock, *Natural strategies for cancer patients*. Kensington Books, 2003.
- [66] R. B. Altman, D. Flockhart, and D. B. Goldstein, *Principles of pharmacogenetics and pharmacogenomics*. Cambridge University Press, 2012.
- [67] W. W. Weber, "Toxicogenomics: History and current applications," *Oncology*, vol. 25, p. 40, 2004.
- [68] G. Orphanides, "Toxicogenomics: challenges and opportunities," *Toxicology letters*, vol. 140, pp. 145–148, 2003.
- [69] L. Brouwers, M. Iskar, G. Zeller, V. Van Noort, and P. Bork, "Network neighbors of drug targets contribute to drug side-effect similarity," *PloS one*, vol. 6, no. 7, p. e22187, 2011.
- [70] M. Yang, X. Wang, and M. Y. Kiang, "Identification of consumer adverse drug reaction messages on social media.," in *PACIS*, p. 193, 2013.
- [71] C. C. Yang, L. Jiang, H. Yang, and X. Tang, "Detecting signals of adverse drug reactions from health consumer contributed content in social media," in *Proceedings of ACM SIGKDD workshop on health informatics*, 2012.
- [72] H. Sampathkumar, B. Luo, and X.-w. Chen, "Mining adverse drug side-effects from online medical forums," in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pp. 150–150, IEEE, 2012.

- [73] J. Liu, A. Li, and S. Seneff, “Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs,” in *Proceedings of First International Conference on Advances in Information Mining and Management (IMMM), Barcelona, Spain*, pp. 23–29, Citeseer, 2011.
- [74] J. Parker, Y. Wei, A. Yates, O. Frieder, and N. Goharian, “A Framework for Detecting Public Health Trends with Twitter,” in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 556–563, ACM, 2013.
- [75] P. Avillach, J.-C. Dufour, G. Diallo, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifirò, A. Fourrier-Réglat, A. Pariente, *et al.*, “Design and Validation of An Automated Method to Detect Known Adverse Drug Reactions in MEDLINE: A Contribution from the EU–ADR Project,” *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 446–452, 2013.
- [76] N. Elhadad, S. Pradhan, W. Chapman, S. Manandhar, and G. Savova, “Semeval-2015 task 14: Analysis of clinical text,” in *Proc of Workshop on Semantic Evaluation. Association for Computational Linguistics*, pp. 303–10, 2015.
- [77] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova, “Semeval-2014 task 7: Analysis of clinical text,” *SemEval*, vol. 199, no. 99, p. 54, 2014.
- [78] I. Segura Bedmar, P. Martínez, and M. Herrero Zazo, “Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013),” in *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, 2013.
- [79] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/va challenge on concepts, assertions, and relations in clinical text,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

- [80] D. Benikova, C. Biemann, M. Kisselew, and S. Padó, “Germeval 2014 named entity recognition shared task: Companion paper,” *Organization*, vol. 7, p. 281, 2014.
- [81] A. R. Aronson, “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap program,” in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.
- [82] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances,” *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.
- [83] A. T. McCray, A. Burgun, and O. Bodenreider, “Aggregating umls semantic types for reducing conceptual complexity,” *Studies in health technology and informatics*, vol. 84, no. 0 1, p. 216, 2001.
- [84] O. Bodenreider and A. T. McCray, “Exploring semantic groups through visual approaches,” *Journal of biomedical informatics*, vol. 36, no. 6, pp. 414–432, 2003.
- [85] A. T. McCray, “An upper-level ontology for the biomedical domain,” *Comparative and Functional Genomics*, vol. 4, no. 1, pp. 80–84, 2003.
- [86] S. Soderland, B. Roof, B. Qin, S. Xu, O. Etzioni, *et al.*, “Adapting open information extraction to domain-specific relations,” *AI magazine*, vol. 31, no. 3, pp. 93–102, 2010.
- [87] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, “Open information extraction from the web,” *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.
- [88] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam, “Open Information Extraction: The Second Generation,” in *IJCAI*, vol. 11, pp. 3–10, 2011.
- [89] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open Information Extraction for the Web,” in *IJCAI*, vol. 7, pp. 2670–2676, 2007.

- [90] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland, “Textrunner: Open Information Extraction on the Web,” in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 25–26, Association for Computational Linguistics, 2007.
- [91] M. Schmitz, R. Bart, S. Soderland, O. Etzioni, *et al.*, “Open Language Learning for Information Extraction,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534, Association for Computational Linguistics, 2012.
- [92] G. Angeli, M. J. Premkumar, and C. D. Manning, “Leveraging Linguistic Structure For Open Domain Information Extraction,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pp. 26–31, 2015.
- [93] J. Xiao, J. Su, G.-d. Zhou, and C. Tan, “Protein-protein interaction extraction: a supervised learning approach,” in *Proc Symp on Semantic Mining in Biomedicine*, pp. 51–59, 2005.
- [94] J. Jiang and C. Zhai, “A systematic exploration of the feature space for relation extraction,” in *HLT-NAACL*, pp. 113–120, 2007.
- [95] J. Li, Z. Zhang, X. Li, and H. Chen, “Kernel-based learning for biomedical relation extraction,” *Journal of the Association for Information Science and Technology*, vol. 59, no. 5, pp. 756–769, 2008.
- [96] J. Jiang, “Information extraction from text,” *Mining text data*, pp. 11–41, 2012.
- [97] I. Segura-Bedmar, P. Martínez, R. Revert, and J. Moreno-Schneider, “Exploring spanish health social media for detecting drug effects,” *BMC medical informatics and decision making*, vol. 15, no. 2, p. S6, 2015.

- [98] S. Taewijit and T. Theeramunkong, “Exploring the distributional semantic relation for adr and therapeutic indication identification in emr,” in *Pacific Rim International Conference on Artificial Intelligence*, pp. 3–15, Springer, 2016.
- [99] Y. Peng, C.-H. Wei, and Z. Lu, “Improving chemical disease relation extraction with rich features and weakly labeled data,” *Journal of Cheminformatics*, vol. 8, no. 1, p. 53, 2016.
- [100] W. M. centre, “Medicines: safety of medicines adverse drug reactions,” vol. Fact sheet N293, 2008.
- [101] M. Liu, E. R. McPeck Hinz, M. E. Matheny, J. C. Denny, J. S. Schildcrout, R. A. Miller, and H. Xu, “Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records,” *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 420–426, 2012.
- [102] E. Roitmann, R. Eriksson, and S. Brunak, “Patient stratification and identification of adverse event correlations in the space of 1190 drug related adverse events,” *Frontiers in physiology*, vol. 5, 2014.
- [103] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman, “Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions,” *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 413–419, 2012.
- [104] M. Y. Park, D. Yoon, K. Lee, S. Y. Kang, I. Park, S.-H. Lee, W. Kim, H. J. Kam, Y.-H. Lee, J. H. Kim, *et al.*, “A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database,” *Pharmacoepidemiology and drug safety*, vol. 20, no. 6, pp. 598–607, 2011.
- [105] S. Skentzos, M. Shubina, J. Plutzky, and A. Turchin, “Structured vs. unstructured: factors affecting adverse drug reaction documentation in an emr repository,” in *AMIA Annual Symposium Proceedings*, vol. 2011, p. 1270, American Medical Informatics Association, 2011.

- [106] Y. Ji, H. Ying, P. Dews, J. Tran, A. Mansour, R. E. Miller, and R. M. Massanari, "An exclusive causal-leverage measure for detecting adverse drug reactions from electronic medical records," in *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American*, pp. 1–6, IEEE, 2011.
- [107] E. Iqbal, R. Mallah, R. G. Jackson, M. Ball, Z. M. Ibrahim, M. Broadbent, O. Dzahini, R. Stewart, C. Johnston, and R. J. Dobson, "Identification of adverse drug events from free text electronic patient records and information in a large mental health case register," *PloS one*, vol. 10, no. 8, p. e0134208, 2015.
- [108] P. L. Peissig, V. S. Costa, M. D. Caldwell, C. Rottschait, R. L. Berg, E. A. Mendonca, and D. Page, "Relational machine learning for electronic health record-driven phenotyping," *Journal of biomedical informatics*, vol. 52, pp. 260–270, 2014.
- [109] Y. Liu, P. LePendur, S. Iyer, and N. H. Shah, "Using temporal patterns in medical records to discern adverse drug events from indications," *AMIA Summits on Translational Science proceedings*, vol. 2012, p. 47, 2012.
- [110] I. Karlsson, J. Zhao, L. Asker, and H. Boström, "Predicting adverse drug events by analyzing electronic patient records," in *Artificial Intelligence in Medicine*, pp. 125–129, Springer, 2013.
- [111] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "Gate: an architecture for development of robust hlt applications," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 168–175, Association for Computational Linguistics, 2002.
- [112] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe, "Extraction of adverse drug effects from clinical records," *Stud Health Technol Inform*, vol. 160, no. Pt 1, pp. 739–43, 2010.
- [113] S. Sohn, J.-P. A. Kocher, C. G. Chute, and G. K. Savova, "Drug side effect extraction from clinical narratives of psychiatry and psychology patients," *Jour-*

- nal of the American Medical Informatics Association*, vol. 18, no. Supplement 1, pp. i144–i149, 2011.
- [114] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, “Identifying adverse drug event information in clinical notes with distributional semantic representations of context,” *Journal of biomedical informatics*, vol. 57, pp. 333–349, 2015.
  - [115] A. Casillas, A. Pérez, M. Oronoz, K. Gojenola, and S. Santiso, “Learning to extract adverse drug reaction events from electronic health records in spanish,” *Expert Systems with Applications*, vol. 61, pp. 235–245, 2016.
  - [116] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13).
  - [117] E. Stammatatos, N. Fakotakis, and G. Kokkinakis, “Automatic Eextraction of Rules for Sentence Boundary Disambiguation,” in *Proceedings of the Workshop on Machine Learning in Human Language Technology*, pp. 88–92, 1999.
  - [118] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.
  - [119] M. Kreuzthaler and S. Schulz, “Detection of Sentence Boundaries and Abbreviations in Clinical Narratives,” *BMC medical informatics and decision making*, vol. 15, no. 2, p. S4, 2015.
  - [120] Z. Apalla, G. Karakatsanis, M. Papageorgiou, C. Kastoridou, and G. Chaide-menos, “A case of atrophoderma vermiculatum responding to systemic isotretinoin,” *Journal of Dermatological case reports*, vol. 3, no. 4, p. 62, 2009.

- [121] F. S. van Dijk, H. Brittain, R. Boerma, M. L. Robert, and J. M. Cobben, “Atrophoderma vermiculatum: A cutaneous feature of loeys-dietz syndrome,” *Jama dermatology*, vol. 151, no. 6, pp. 675–677, 2015.
- [122] F. Wang, P. Zhang, N. Cao, J. Hu, and R. Sorrentino, “Exploring the associations between drug side-effects and therapeutic indications,” *Journal of biomedical informatics*, vol. 51, pp. 15–23, 2014.
- [123] I. Segura-Bedmar, S. De La Pena, and P. Martinez, “Extracting drug indications and adverse drug reactions from spanish health social media,” in *ACL*, vol. 2014, p. 98, 2014.
- [124] R. Xu and Q. Wang, “Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing,” *BMC bioinformatics*, vol. 14, no. 1, p. 181, 2013.
- [125] R. Xu and Q. Wang, “Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature,” *Journal of biomedical informatics*, vol. 51, pp. 191–199, 2014.
- [126] S. Gupta and C. D. Manning, “Improved pattern learning for bootstrapped entity extraction,” in *CoNLL*, pp. 98–108, 2014.
- [127] E. Agichtein and L. Gravano, “Snowball: Extracting relations from large plain-text collections,” in *Proceedings of the fifth ACM conference on Digital libraries*, pp. 85–94, ACM, 2000.
- [128] L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-based information extraction is dead! long live rule-based information extraction systems!,” in *EMNLP*, no. October, pp. 827–832, 2013.
- [129] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011, Association for Computational Linguistics, 2009.

- [130] W. Zheng and C. Blake, “Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles,” *Journal of biomedical informatics*, vol. 57, pp. 134–144, 2015.
- [131] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open Information Extraction from the Web,” in *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, (San Francisco, CA, USA), pp. 2670–2676, Morgan Kaufmann Publishers Inc., 2007.
- [132] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, “Open Information Extraction from the Web,” *Commun. ACM*, vol. 51, pp. 68–74, Dec. 2008.
- [133] S. Taewijit, T. Theeramunkong, and M. Ikeda, “Distant supervision with transductive learning for adverse drug reaction identification from electronic medical records,” *Journal of Healthcare Engineering*, vol. 2017, 2017.
- [134] R. Feldman, O. Netzer, A. Peretz, and B. Rosenfeld, “Utilizing text mining on online medical forums to predict label change due to adverse drug reactions,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1779–1788, ACM, 2015.
- [135] A. Nikfarjam, A. Sarker, K. OConnor, R. Ginn, and G. Gonzalez, “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features,” *Journal of the American Medical Informatics Association*, p. ocu041, 2015.
- [136] S. Eyheramendy, D. D. Lewis, and D. Madigan, “On the naive bayes model for text categorization,” 2003.
- [137] O. Maron and T. Lozano-Pérez, “A framework for multiple-instance learning,” in *Advances in neural information processing systems*, pp. 570–576, 1998.
- [138] J. Liu, S. Zhao, and X. Zhang, “An ensemble method for extracting adverse drug events from social media,” *Artificial intelligence in medicine*, vol. 70, pp. 62–76, 2016.

- [139] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Knowledge-based extraction of adverse drug events from biomedical text," *BMC bioinformatics*, vol. 15, no. 1, p. 64, 2014.
- [140] X. Liu and H. Chen, "A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports," *Journal of biomedical informatics*, vol. 58, pp. 268–279, 2015.
- [141] N. Kasabov and S. Pang, "Transductive support vector machines and applications in bioinformatics for promoter recognition," in *Neural networks and signal processing, 2003. proceedings of the 2003 international conference on*, vol. 1, pp. 1–6, IEEE, 2003.
- [142] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.
- [143] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

# Publications

## Peer-Reviewed International Journal

- [1] Siriwon Taewijit, Thanaruk Theeramunkong, Mitsuru Ikeda: “Distant Supervision with Transductive Learning for Adverse Drug Reaction Identification from Electronic Medical Records”, Journal of Healthcare Engineering, (2017):1-22, 2017. (*Published*)
- [2] Tu Bao Ho, Ly Le, Dang Tran Thai, Siriwon Taewijit: “Data-driven Approach to Detect and Predict Adverse Drug Reactions”, Journal of Current Pharmaceutical Design, 22(23):3498-526, 2016

## Peer-Reviewed International Conferences

- [3] Siriwon Taewijit, Thanaruk Theeramunkong: “Distant Supervision with Transductive Learning for Adverse Drug Reaction Identification from Electronic Medical Records”, Workshop on Big data analytics-as-a-Service: Architecture, Algorithms, and Application in Health Informatics, 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2017), August 13-17, 2017.
- [4] Siriwon Taewijit, Thanaruk Theeramunkong: “Exploring the Distributional Semantic Relation for ADR and Therapeutic Indication Identification in EMR”, In Pacific Rim International Conference on Artificial Intelligence, pp. 3-15, August 22-26, 2016. (*The best presentation award*)

- [5] Siriwon Taewijit, Thanaruk Theeramunkong: “A Probabilistic Approach for Customer Repurchase Intention and Association-Based Product Network Analysis: An Empirical Study of Fast Food Market”, In Proceedings of Asian Conference on Information Systems 2014, pp. 339-346, December 1- 3, 2014.
- [6] Siriwon Taewijit, Tu Bao Ho, Thanaruk Theeramunkong: “A Review on Data Mining Approach for Adverse Drug Reaction Research”, In Proceedings of the Second Asian Conference on Information Systems 2013, pp. 103-110, October 31- November 2, 2013.

## **Peer-Reviewed International Conferences (Poster)**

- [7] Siriwon Taewijit, Thanaruk Theeramunkong: “Exploring the Distributional Semantic Relation for ADR and Therapeutic Indication Identification in EMR”, The 11th International Symposium in Science and Technology at Kansai University, July 26-28, 2016.
- [8] Siriwon Taewijit, Tu Bao Ho: “A Two-Stage Approach for Multivariate Infrequent Adverse Drug Events Analysis in Electronic Medical Records”, In Proceedings of the Fifth Annual Translational Bioinformatics Conference 2015, pp. 46, November 7-9, 2015.

## **Book Chapter**

- [9] Tu Bao Ho, Siriwon Taewijit, Quang Bach Ho, Hieu Chi Dam: “Progressive Trends in Knowledge and System-based Science for Service Innovation: (Chapter 7) Big Data and Service Science”, IGI Global, pp. 128-144, 2013