Relation extraction methods for biomedical literature

Bui, Q.C.

Publication date
2012

Chapter 1. Introduction

1.1 Motivation

Biomedical data is crucial for research in life sciences such as systems biology and medicine. For instance, systems biology needs data to understand properties of protein-protein interactions (PPIs) [1]. Interactions data, e.g. data stored in PPIs databases, are used to validate the reliability of large-scale experimental PPI datasets [2, 3]. Networks generated from these interactions are useful for multiple purposes, e.g. for predicting new PPIs and for finding novel associations between genes and diseases[4–6]. Furthermore, experimental data are also systematically linked with the results of different studies derived from published literature to obtain a better understanding of the biological systems, and to find interesting associations among disparate facts, leading to the discovery of new or unsuspected knowledge [7, 8]. In medicine, up-to-date drug resistant information is vital to have better treatment for many diseases [9, 10]. *In silico*, biomedical data are used to build computational models to simulate how viruses, for example human immunodeficiency virus (HIV), interact with the human immune system in order to find an optimal drug regiment for individual treatment [11–13]. HIV epidemic data is used to study how these viruses are spread due to social interactions in order to have a better prevention strategy [14]. Above all, to facilitate these studies, data need to be in structured form for easy access and for obtaining evidence [15–18]. A typical example that shows the role of biomedical data in identifying diseases is illustrated in Figure 1.1.

Vast amount of biomedical data are available in unstructured form through scientific publications. Together with the development of high-throughput experimental techniques and computational models, this amount of data is being generated with an exponential rate [7, 10, 17]. Traditional search engine such as Google or specialized information retrieval tools such as PubMed provide modest help. With a few keywords, PubMed or Google can return thousands of relevance documents, but the users still need to read all those returned documents to find the data they need. Although new requirements for publishing biomedical results have been proposed such as nanopublication [19] or annotated digital abstracts [20], it is becoming more and more difficult to discover knowledge or generate scientific hypotheses without the use of data extraction techniques [16, 17]. At the same time, the number of biological databases and their entries, e.g. BIND [21], BioGrid [22], and HPRD [23], grow steadily but most of information is added by database curators and does come directly not from the original research. To build such PPIs databases, literature curators have to read all documents related to the field. This is a time consuming and laborious task. Furthermore, a recent study by [24] shows that the quality of literature-curated PPI databases is relatively low and can only cover a small fraction of data actually discovered. One of the reasons is that new discovered data are published in a wide range of scientific journals which spread into many

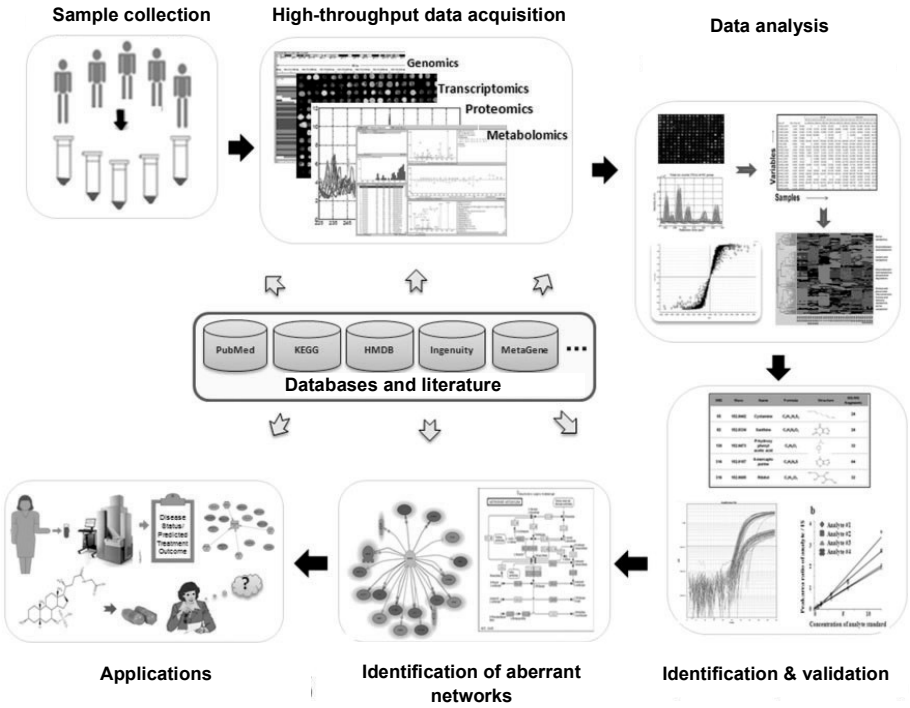

Figure 1.1 The central role of biomedical data in life sciences (adapted from [16])

disciplines and have a weak link to each other. Therefore many PPIs are overlooked by the curators. For these reasons, there is increasing interest in techniques that can automatically extract relations between entities such as PPIs from biomedical text and present the distilled knowledge to users in concise and structured form [25]. Studying these relation extraction methods is the main focus of this thesis.

There are many challenges to extract relations from biomedical text which requires new and different methods compared to the existing approaches being used for general text. First, the high ambiguity of vocabulary, long and complex sentences in biomedical text cause performance of natural language processing (NLP) tools, which trained on general English corpora, to drop considerably [26, 27]. Therefore the performance of many relation extraction methods that work well for newswire text degrades significantly when applied to text in the biomedical domain. Second, the lack of standard gene and protein names and their synonyms make the recognition of name entities (NER) mentioned in biomedical text, which is a prerequisite step of relation extraction, a difficult task [3, 28]. Furthermore, the high degree of variation in biomedical terminologies also contributes to this problem, which then degrades the overall performance of the extraction systems [25]. Finally, the availability of high quality annotated corpora is scarce since they are expensive

and time-consuming to produce. Such corpora are important to train NLP tools as well as machine learning (ML) algorithms for extracting relations in biomedical text [29–32]. Due to these challenges, extracting relations from biomedical text has been an active research field during the last decade.

Although many approaches have been proposed, extracting relations from biomedical text remains a big issue due to, among others, the quality of the extracted relations, performance time (speed), and the type of relations being extracted [28, 33–37]. First, the performance of extraction systems, which is measured in terms of precision, needs to be improved to satisfy the demand of aforementioned tasks such as building high quality biological databases. Second, most of the proposed systems require a significant performance time when applied for large scale extraction. Therefore, these systems are not ready for real time application. Third, existing approaches have mainly focused on extracting PPIs, and recently on *biomedical events*; many relation types are still untouched.

1.2 Research questions

The aim of this thesis is to study methods for relation extraction from biomedical text. In particular, we focus on extracting three types of relations, namely causal relations on HIV drug resistance, protein-protein interactions, and *biomedical events*. Furthermore, these relation extraction methods are investigated under three different scenarios that are commonly encountered in this research field: no training data is available, training data is available but relation types are missing, and full training data are available, respectively. The research questions that we address in this thesis are as follows:

1. How is syntactic information used for different relation extraction tasks? Syntactic information has been used in most relation extraction systems. The use of syntactic information depends on the levels of NLP analysis ranging from part-of-speech (POS) to deep parsing. In this thesis we study the use of syntactic information for three relation extraction methods in accordance with the scenarios above. This research question is addressed in Chapter 3, 4 and 5.

2. What is the role of machine learning to relation extraction task? ML methods play an important role in relation extraction systems. The use of ML methods depends on the availability of training data as well as the properties of data. In this study we demonstrate that ML methods can effectively be used to leverage the performance of existing relation extraction tasks. The role of machine learning to relations extraction task is discussed in Chapter 4 and 5.

3. Which factors contribute to the performance time of a relation extraction system? When the system is applied to large-scale extractions then

computational resources required to train and run the system should be taken into account. There are many factors contribute to the performance time of a relation extraction system such as which NLP tools are used to analyze input text, how many features are used for the ML classifier. Here we show that by applying a data partition strategy for input texts, the performance time of the ML-based systems can be boosted significantly. This research question is addressed in Chapter 4 and 5.

The detailed answers to each research question are given in chapter 6.

1.3 Outline of the thesis

Following the aim of relation extraction: "to distil knowledge and present it in a concise form to users" [6], this thesis is written in a concise form. When details are needed, readers are directed to the specific references. This thesis consists of the following chapters:

Chapter 2 provides background for relation extraction methods in biomedical text. We start with an overview of a typical relation extraction system such as its work-flow and which tools and techniques are commonly used. We then discuss the role of NLP and ML tools for relation extraction tasks. Finally we characterize main techniques and present their state-of-the-art results on relation extraction tasks.

Chapter 3 presents a novel method to extract causal relations on HIV drug resistant based on grammatical rules. We show how these rules are formulated and how to combine the extracted relations. First, we review existing methods to extract binary relations. We then present our method and discuss its results and some possible ways to improve the performance of our method.

Chapter 4 introduces a new method to extract PPIs from biomedical text. We start with an introduction to the existing PPIs extraction approaches. We then describe our method and show how to use a ML classifier as a filter to relation extraction task. Finally, we discuss the evaluation results on five PPI corpora and compare our results with the results of the other systems.

Chapter 5 focuses on methods to extract *biomedical events* from text. We begin with an overview of the *biological events* and existing methods used to extract these events. We then introduce our approach to automatically learn rules from training data and how to apply these rules to new text.

Chapter 6 answers the research questions raised in Chapter 1 and summarizes this thesis with overall discussion and conclusion.