Information Extraction in Biomedical Literature

Min Song
Drexel University, USA

Il-Yeol Song
Drexel University, USA

Xiaohua Hu
Drexel University, USA

Hyoil Han
Drexel University, USA

INTRODUCTION

Information extraction (IE) technology has been defined and developed through the US DARPA Message Understanding Conferences (MUCs). IE refers to the identification of instances of particular events and relationships from unstructured natural language text documents into a structured representation or relational table in databases. It has proved successful at extracting information from various domains, such as the Latin American terrorism, to identify patterns related to terrorist activities (MUC-4). Another domain, in the light of exploiting the wealth of natural language documents, is to extract the knowledge or information from these unstructured plain-text files into a structured or relational form. This form is suitable for sophisticated query processing, for integration with relational databases, and for data mining. Thus, IE is a crucial step for fully making text files more easily accessible.

BACKGROUND

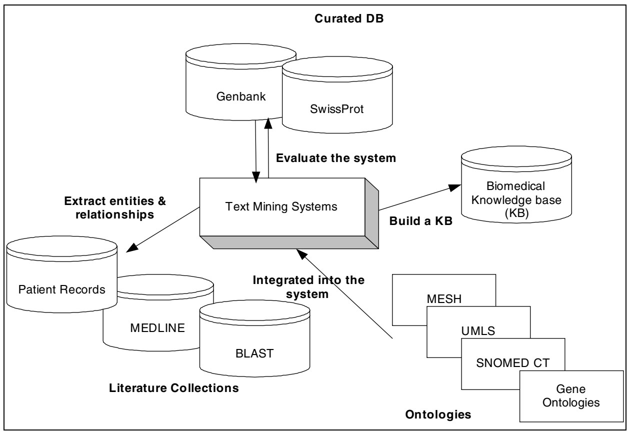
The advent of large volumes of text databases and search engines have made them readily available to domain experts and have significantly accelerated research on bioinformatics. With the size of a digital library commonly exceeding millions of documents, rapidly increasing, and covering a wide range of topics, efficient and automatic extraction of meaningful data and relations has become a challenging issue. To tackle this issue, rigorous studies have been carried out recently to apply IE to biomedical data. Such research efforts began to be called biomedical literature mining or text mining in bioinformatics (de Bruijn & Martin, 2002; Hirschman et al., 2002; Shatkay & Feldman, 2003). In this article, we review recent advances in applying IE techniques to biomedical literature.

MAIN THRUST

This article attempts to synthesize the works that have been done in the field. Taxonomy helps us understand the accomplishments and challenges in this emerging field. In this article, we use the following set of criteria to classify the biomedical literature mining related studies:

1. What are the target objects that are to be extracted?
2. What techniques are used to extract the target objects from the biomedical literature?
3. How are the techniques or systems evaluated?

Figure 1. Shows the overview of a typical biomedical literature mining system.



4) From what data sources are the target objects extracted?

Target Objects

In terms of what is to be extracted by the systems, most studies can be broken into the following two major areas: (1) named entity extraction such as proteins or genes; and (2) relation extraction, such as relationships between proteins. Most of these studies adopt information extraction techniques using curated lexicon or natural language processing for identifying relevant tokens such as words or phrases in text (Shatkay & Feldman, 2003).

In the area of named entity extraction, Proux et al. (2000) use single word names only with selected test set from 1,200 sentences coming from Flybase. Collier, et al. (2000) adopt Hidden Markov Models (HMMs) for 10 test classes with small training and test sets. Krauthammer et al. (2000) use BLAST database with letters encoded as 4-tuples of DNA. Demetriou and Gaizuaskas (2002) pipeline the mining processes, including hand-crafted components and machine learning components. For the study, they use large lexicon and morphology components. Narayanaswamy et al. (2003) use a part of speech (POS) tagger for tagging the parsed MEDLINE abstracts. Although Narayanaswamy and his colleagues (2003) implement an automatic protein name detection system, the number of words used is 302, and, thus, it is difficult to see the quality of their system, since the size of the test data is too small. Yamamoto, et al. (2003) use morphological analysis techniques for preprocessing protein name tagging and apply support vector machine (SVM) for extracting protein names. They found that increasing training data from 390 abstracts to 1,600 abstracts improved F-value performance from 70% to 75%. Lee et al. (2003) combined an SVM and dictionary lookup for named entity recognition. Their approach is based on two phases: the first phase is identification of each entity with an SVM classifier, and the second phase is post-processing to correct the errors by the SVM with a simple dictionary lookup. Bunescu, et al. (2004) studied protein name identification and protein-protein interaction. Among several approaches used in their study, the main two ways are one using POS tagging and the other using the generalized dictionary-based tagging. Their dictionary-based tagging presents higher F-value. Table 1 summarizes the works in the areas of named entity extraction in biomedical literature.

The second target object type of biomedical literature extraction is relation extraction. Leek (1997) applies HMM techniques to identify gene names and chromosomes through heuristics. Blaschke et al. (1999) extract protein-protein interactions based on co-occurrence of the form "… p1…I1… p2" within a sentence, where p1, p2 are proteins, and I1 is an interaction term. Protein names and interaction terms (e.g., activate, bind, inhibit) are provided as a dictionary. Proux (2000) extracts an interact relation for the gene entity from Flybase database. Pustejovsky (2002) extracts an inhibit relation for the gene entity from MEDLINE. Jenssen, et al. (2001) extract a gene-gene relations based on co-occurrence of the form "… g1…g2…" within a MEDLINE abstracts, where g1 and g2 are gene names. Gene names are provided as a dictionary, harvested from HUGO, LocusLink, and other sources. Although their study uses 13,712 named human genes and millions of MEDLINE abstracts, no extensive quantitative results are reported and analyzed. Friedman, et al. (2001) extract a pathway relation for various biological entities from a variety of articles. In their work, the precision of the experiments is high (from 79-96%). However, the recalls are relatively low (from 21-72%). Bunescu et al. (2004) conducted protein/protein interaction identification with several learning methods, such as pattern matching rule induction (RAPIER), boosted wrapper induction (BWI), and extraction using longest common subsequences (ELCS). ELCS automatically learns rules for extracting protein interactions using a bottom-up

Table 1. A summary of works in biomedical entity extraction

Author	Named Entities	Database	No. of Words	Learning Methods	F Value
Collier, et al. (2000)	Proteins and DNA	MEDLINE	30,000	HMM	73
Krauthammer, et al. (2000)	Gene and Protein	Review articles	5,000	Character sequence mapping	75
Demetriou and Gaizauskas (2002)	Protein, Species, and 10 more	MEDLINE	30,000	PASTA template filing	83
Narayanaswamy (2003)	Protein	MEDLINE	302	Hand-crafted rules and co-occurrence	75.86
Yamamoto, et al. (2003)	Protein	GENIA	1,600 abstracts	BaseNP recognition	75
Lee, et al. (2003)	Protein DNA RNA	GENIA	10,000	SVM	77
Bunescu (2004)	Protein	MEDLINE	5,206	RAPIER, BWI, TBL, k-NN , SVMs, MaxEnt	57.86