Chapter 2

INFORMATION EXTRACTION FROM TEXT

Jing Jiang
Singapore Management University
jingjiang@smu.edu.sg

Abstract Information extraction is the task of finding structured information from unstructured or semi-structured text. It is an important task in text mining and has been extensively studied in various research communities including natural language processing, information retrieval and Web mining. It has a wide range of applications in domains such as biomedical literature mining and business intelligence. Two fundamental tasks of information extraction are named entity recognition and relation extraction. The former refers to finding names of entities such as people, organizations and locations. The latter refers to finding the semantic relations such as `FounderOf` and `HeadquarteredIn` between entities. In this chapter we provide a survey of the major work on named entity recognition and relation extraction in the past few decades, with a focus on work from the natural language processing community.

Keywords: Information extraction, named entity recognition, relation extraction

1. Introduction

Information extraction from text is an important task in text mining. The general goal of information extraction is to discover structured information from unstructured or semi-structured text. For example, given the following English sentence,

In 1998, Larry Page and Sergey Brin founded Google Inc.

we can extract the following information,

`FounderOf(`*Larry Page*`,` *Google Inc.*`)`,
`FounderOf(`*Sergey Brin*`,` *Google Inc.*`)`,
`FoundedIn(`*Google Inc.*`,` *1998*`)`.

C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_2,

Such information can be directly presented to an end user, or more commonly, it can be used by other computer systems such as search engines and database management systems to provide better services to end users.

Information extraction has applications in a wide range of domains. The specific type and structure of the information to be extracted depend on the need of the particular application. We give some example applications of information extraction below:

- Biomedical researchers often need to sift through a large amount of scientific publications to look for discoveries related to particular genes, proteins or other biomedical entities. To assist this effort, simple search based on keyword matching may not suffice because biomedical entities often have synonyms and ambiguous names, making it hard to accurately retrieve relevant documents. A critical task in biomedical literature mining is therefore to automatically identify mentions of biomedical entities from text and to link them to their corresponding entries in existing knowledge bases such as the FlyBase.

- Financial professionals often need to seek specific pieces of information from news articles to help their day-to-day decision making. For example, a finance company may need to know all the company takeovers that take place during a certain time span and the details of each acquisition. Automatically finding such information from text requires standard information extraction technologies such as named entity recognition and relation extraction.

- Intelligence analysts review large amounts of text to search for information such as people involved in terrorism events, the weapons used and the targets of the attacks. While information retrieval technologies can be used to quickly locate documents that describe terrorism events, information extraction technologies are needed to further pinpoint the specific information units within these documents.

- With the fast growth of the Web, search engines have become an integral part of people's daily lives, and users' search behaviors are much better understood now. Search based on bag-of-word representation of documents can no longer provide satisfactory results. More advanced search problems such as entity search, structured search and question answering can provide users with better search experience. To facilitate these search capabilities, information ex-