IRM PRESS
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB9195

Chapter II

Information Extraction from Free-Text Business Documents

Witold Abramowicz
The Poznan University of Economics, Poland

Jakub Piskorski
German Research Center for Artificial Intelligence in Saarbruecken, Germany

ABSTRACT

The objective of this chapter is an investigation of the applicability of information extraction techniques in real-world business applications dealing with textual data since business relevant data is mainly transmitted through free-text documents. In particular, we give an overview of the information extraction task, designing information extraction systems and some examples of existing information extraction systems applied in the financial, insurance and legal domains. Furthermore, we demonstrate the enormous indexing potential of lightweight linguistic text processing techniques applied in information extraction systems and other closely related fields of information technology which concern processing vast amounts of textual data.

INTRODUCTION

Nowadays, knowledge relevant to business of any kind is mainly transmitted through free-text documents: the World Wide Web, newswire feeds, corporate reports, government documents, litigation records, etc. One of the most difficult issues concerning applying search technology for retrieving relevant information from textual data

collections is the process of converting such data into a shape for searching. Information retrieval (IR) systems using conventional indexing techniques applied even to a homogenous collection of text documents fall far from obtaining optimal recall and precision simultaneously. Since structured data is obviously easier to search, an ever-growing need for effective and intelligent techniques for analyzing free-text documents and building expressive representation of their content in the form of structured data can be observed.

Recent trends in information technology such as Information Extraction (IE) provide dramatic improvements in converting the overflow of raw textual information into valuable and structured data, which could be further used as input for data mining engines for discovering more complex patterns in textual data collections. The task of IE is to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information, where the domain consists of a corpus of texts together with a clearly specified information need. Due to the specific phenomena and complexity of natural language, this is a non-trivial task. However, recent advances in Natural Language Processing (NLP) concerning new robust, efficient, high coverage shallow processing techniques for analyzing free text have contributed to the size in the deployment of IE techniques in business information systems.

INFORMATION EXTRACTION

Information Extraction Task

The task of IE is to identify instances of a particular pre-specified class of events or relationships and entities in natural language texts, and the extraction of the relevant arguments of the events or relationships (SAIC, 1998). The information to be extracted is pre-specified in user-defined structures called templates (e.g., company information, meetings of important people), each consisting of a number of slots, which must be instantiated by an IE system as it processes the text. The slots are usually filled with: some strings from the text, one of a number of pre-defined values or a reference to other already generated template. One way of thinking about an IE system is in terms of database construction, since an IE system creates a structured representation of selected information drawn from the analyzed text.

In recent years IE technology has progressed quite rapidly, from small-scale systems applicable within very limited domains to useful systems which can perform information extraction from a very broad range of texts. IE technology is now coming to the market and is of great significance to finance companies, banks, publishers and governments. For instance, a financial organization would want to know facts about foundations of international joint-ventures happening in a given time span. The process of extracting such information involves locating the names of companies and finding linguistic relations between them and other relevant entities (e.g., locations and temporal expressions). However, in this particular scenario an IE system requires some specific domain knowledge (understanding the fact that ventures generally involve at least two partners and result in the formation of a new company) in order to merge partial information into an adequate template structure. Generally, IE systems rely to some degree on domain knowledge. Further information such as appointment of key personnel