Chapter 3

Information Extraction from Text and Beyond

Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium

ABSTRACT

In this chapter, the author defines information extraction from text, describes common information extraction tasks, and discusses current information extraction issues being the need to develop technologies that require a minimum of human supervision, to build systems that automatically acquire world knowledge, and to integrate their outputs into advanced information extraction systems. Current emerging research on extraction of narrative scenarios and complex concepts revives an old dream and opens a way to full natural language understanding.

INTRODUCTION

Information extraction is a natural language processing technique which aims at identifying information in unstructured data. Written and spoken text, pictures, video and audio are all forms of unstructured data. Unstructured does not imply that the data is structurally incoherent (in that case it would simply be nonsense), but rather that its information is encoded in such a way that makes it difficult for computers to immediately interpret it. Information extraction is the process that adds meaning to unstructured, raw data, whether that is text, images, video or audio. Consequently, the data become structured or semi-structured and can be more easily processed by the computer (e.g., in information retrieval, data mining, data visualization, and data summarization). In other words, information extraction presupposes that although the semantic information in a text and its linguistic organization is not immediately computationally transparent, it can nevertheless be retrieved by taking into account surface regularities that reflect its computationally opaque
An information extraction system will use a set of extraction patterns, which are either manually constructed or automatically learned, to take information out of a text and put it in a more structured format. In this chapter we adopt the following definition of information extraction: (Moens 2006 p. 225):

**Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, providing additional aids to access and interpret the unstructured data by information systems.**

During the Message Understanding Conferences (MUC) in the late 1980s and 1990s, there gradually arose a set of typical information extraction tasks (see Grishman & Sundheim 1996; Cunningham 1997). A most popular task probably is **named entity recognition**, i.e., recognizing person names, organizations, locations, date, time, money and percents. These names are often expanded to protein names, product brands, etc. Other tasks are **event extraction**, i.e., recognizing events, their participants and settings, **scenario extraction**, i.e., linking of individual events in a story line, and complex concept extraction, i.e., linking individual events into an abstract concept or issue. Currently, **relation extraction**, where the relationship between two or more entities is defined (e.g., Figure 1), respectively referred to as binary and n-ary relation extraction, is quite popular. Entities, relations or scenarios are semantically typed. Other complementary or more specialized tasks embedded in discourse processing regard **semantic role labeling**, i.e., finding the predicates and arguments in sentences that typically represent the actions, actors and circumstances in a sentence (“who” did “what to “whom” “where” and “when”) (Marquez et al. 2008) and **coreference resolution**, i.e., determining whether two expressions in natural language refer to the same entity, person, time, place, and event in the world (e.g., Ng 2008; Ragman & Ng 2009), **recognition of temporal expressions and relations** (Boguraev et al. 2005; Kolomiyets & Moens 2009; Yoshikawa et al. 2009), **recognition of spatial expressions and relations** (Katz 2004; Regier et al. 2005), and **recognition of causal relations** (Girju 2002). Apart from these many other information extraction tasks can be defined such as opinion extraction (Pang & Lee 2008; Boiy & Moens 2009), argumentation extraction (Mochales & Moens 2009) and many others.

Typical **extraction units** for an extraction system are word compounds and basic noun phrases, but in some applications it might be opportune to extract other linguistic units, such as verb phrases, temporal markers, subclauses or larger rhetorical structures. Information can be extracted from one clause or from multiple clauses or sentences spanning one or more texts before it is outputted by the system. Consider the example that an information question wants to retrieve event information about assassinations, it might be that the name of the person assassinated and the time and place of the event is named in a first sentence of a news article, but that the name of the assassin and his method are mentioned in some sentences further in the discourse.

Information extraction is not to be seen as a replacement of the raw data. It only offers additional layers or views of the information. Consequently, information extraction is not a stand-alone application. The extracted information is used for some purpose in a follow-up task. We assign semantic