Chapter 75
Information Extraction in the Medical Domain

Aicha Ghoulam
University of Oran 1, Ahmed Ben Bella, Algeria

Fatiha Barigou
University of Oran 1, Ahmed Ben Bella, Algeria

Ghalem Belalem
University of Oran 1, Ahmed Ben Bella, Algeria

ABSTRACT

Information Extraction (IE) is a natural language processing (NLP) task whose aim is to analyse texts written in natural language to extract structured and useful information such as named entities and semantic relations between them. Information extraction is an important task in a diverse set of applications like bio-medical literature mining, customer care, community websites, personal information management and so on. In this paper, the authors focus only on information extraction from clinical reports. The two most fundamental tasks in information extraction are discussed; namely, named entity recognition task and relation extraction task. The authors give details about the most used rule/pattern-based and machine learning techniques for each task. They also make comparisons between these techniques and summarize the advantages and disadvantages of each one.

1. INTRODUCTION

The amount of information written in natural language and available in electronic format is increasing. Due to their unstructured nature, however, manual analysis of this huge information is challenging and labor intensive. To address these concerns we need new techniques of structured extraction to access useful information. Information Extraction (IE) can relieve some of these problems by offering access to relevant information without requiring the end user of the information to read the text.

As it is mentioned in (Jiang, 2012), extraction of structured information from text dates back to the ‘70s, it started gaining much attention when DARPA (Defense Advanced Research Projects Agency)

DOI: 10.4018/978-1-5225-1759-7.ch075

initiated and funded the Message Understanding Conferences (MUC) in the '90s,. MUCs defined information extraction as filling a predefined template that contains a set of predefined slots like a terrorism template used in MUC-4. Template filling is a complex task and systems developed to fill one template cannot directly work for a different template. In MUC-6, a number of template-independent subtasks of information extraction were defined; these include named entity recognition, and relation extraction.

Early information extraction systems like the ones that participated in the MUCs were rule-based with manually coded rules. They use linguistic extraction patterns developed by humans to match text and locate information units. They can achieve good performance on a specific target domain, but it is labor intensive to design good extraction rules, and the developed rules are highly domain dependent. Realizing the limitations of these manual developed systems, researchers turned to statistical machine learning approaches. With the decomposition of information extraction systems into components such as named entity recognition, many information extraction subtasks can be transformed into classification problems or sequence labeling, the first one can be solved by standard supervised learning algorithms such as support vector machines and maximum entropy models, and the second one because information extraction involves identifying segments of text that play different roles, it can be solved by hidden Markov models and conditional random fields.

The IE is a research subject that covers many areas like customer care, personal information management, bio-informatics, community web sites. As it is mentioned in (Berrazega, 2012); these applications require IE for searching and responding queries.

To facilitate these search capabilities, information extraction is often needed as a preprocessing step to enrich document representation or to populate a database.

As the volume of medical knowledge double every five years according to some studies as it mentioned in (Ben Abacha & Zweigenbaum, 2011a), and recorded in unstructured formats, development of medical information extraction techniques have gained immense popularity. They include identification of biomedical and/or medical named entities, relations between the entities, or events associated like the one developed in (Zweigenbaum & Tannier, 2013).

Noticeable efforts have been invested in the medical domain. Examples include the work of (Harkema et al., 2005) who applied AMBIT in clinical and biomedical texts to extract key information. Aronson (2001) used MetaMap tool to recognize and categories medical terms.

In this paper, we focus on the two most fundamental tasks in information extraction, namely, named entity recognition and relation extraction in the medical field. We will compare some works using rule/pattern-based and machine learning approaches in term of used corpus, coverage and precision. The remainder of this paper is divided as follows: Section 2 presents the information extraction concept and approaches of extraction. Section 3 introduces information extraction in the medical domain, look at the related work on medical information extraction, and then initiate a comparative study. Finally, section 4 presents our conclusions and perspectives.

2. INFORMATION EXTRACTION

2.1. Definition

Information Extraction has been defined in the literature review by many researchers (Sarawagi, 2007) and (Jiang, 2012). The most common definition is that IE is an automatic process for extracting structured