Chapter 78
Events Automatic Extraction from Arabic Texts

Emna Hkiri
University of Monastir, Tunisia

Souheyl Mallat
University of Monastir, Tunisia

Mounir Zrigui
University of Monastir, Tunisia

ABSTRACT

The event extraction task consists in determining and classifying events within an open-domain text. It is very new for the Arabic language, whereas it attained its maturity for some languages such as English and French. Events extraction was also proved to help Natural Language Processing tasks such as Information Retrieval and Question Answering, text mining, machine translation etc… to obtain a higher performance. In this article, we present an ongoing effort to build a system for event extraction from Arabic texts using Gate platform and other tools.

1. INTRODUCTION

With the increase amount of textual data, there is a need of methodologies for analyzing, and representing these data in the best form. One of the challenges of the textual data analysis is information extraction (IE). IE as defined in the message understanding conferences (MUC), aims to analyze natural language texts and extract useful information from a particular domain or application. IE is used in processing text for Arabic systems such as search engines, clustering, classification, text mining systems Question and answering, information retrieval (Atta-Allah al, 2006), text summary, indexation, and classification(Cimiano et al, 2004), (Cohen et al, 2009).

DOI: 10.4018/978-1-7998-0951-7.ch078

If this search domain is already extensively covered, the event extraction remains a vital and complex task. Event extraction from Arabic text could be beneficial for IE systems in various ways. For example, being able to determine events could improve the performance of news systems (Borsje et al, 2010), monitoring systems (kamijo et al, 2000) risk analysis applications (Capet et al, 2008) and decision making tools (wei et al, 2004).Extracted events are also extensively applied within the medical domain (Cohen et al, 2009), (Yakuji et al, 2001)

In this paper, the input of our information extraction system is a set of unclassified news websites articles written in Arabic language, and the output is a set of events with their attributes. In this context, we adopt the definition of the ACE (Automatic Content Extraction) (George, 2004) model that defines the event as an action, a process in which participants are connected. We describe in this paper the different steps followed to develop our event extraction system from Arabic texts.

The rest of the paper is organized as follows: Section (2) presents Arabic language and details its unique features. Section (3) deals with the definition of the event and present some related works. In section (3), we present our approach for the automatic extraction of events. The implementation of our approach in the Gate platform is described in section (4). In section (5), we evaluate the system to demonstrate its capabilities. Finally, we conclude our work with some perspectives.

2. ARABIC LANGUAGE

In Arabic processing domain, the research started in the 1970, even before editing texts problems are fully solved. Early studies focused mainly on lexicons. For ten years, the Web internationalization and the proliferation of media in Arabic demonstrated the usefulness of a large number of potential applications of the Arabic NLP. Therefore, researches have begun to address issues more varied as syntax, automatic translation, automatic indexing of documents, information retrieval, etc. (Farber; 2008).

In what follows, we present orthographic and morphological systems of the Arabic language and some problems of its automatic processing as the lack of free resources, lack of vowels and the agglutination of words. Those are the main issues that characterize Arabic and strongly contribute to the delay of its automatic processing. For morphological analysis, the absence of vowels adds additional ambiguity of Arabic words. As for the agglutination, it makes it more difficult to identify the segments that make up these words.

2.1. Features of Arabic Language

Arabic language is considered difficult to control in the NLP. It is a semitic language: it is written from right to left and its alphabet is an Abjad. It is composed of consonants and long vowels "ي", "ا" and "و".

In addition, Arabic writing is unicameral: capital letters and lowercase letters do not exist. It is a semi cursive language that most of its letters are attached to each other, their spellings differ depending on whether they are preceded or / and followed by other letters or are isolated. Only the letters "ا" and "و, د, ر" are never attach to the next letter.

The Arabic lexicon is composed of three main classes: nouns, verbs and particles. The class of particles is extended to include grammatical morphemes. This extension leads to a reorganization into four classes: verbs, nouns, pronouns and tools words and (Khoja et al, 2001):