King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Automatic extraction of ontological relations from Arabic text

Mohammed G.H. Al Zamil [*], Qasem Al-Radaideh

Department of Computer Information Systems, Yarmouk University, Irbed, Jordan

Available online 28 September 2014

KEYWORDS

Arabic ontology;
Lexical syntactic patterns;
Automatic extraction of relationships

Abstract Automatic extraction of semantic relationships among Arabic concepts to formulate ontology models is crucial for providing rich semantic metadata. Due to the annual increase of Arabic content on the Internet, the need for specialized tools to analyze and understand Arabic text has emerged. This research proposes a methodology that extracts ontological relationships. The goals of this research are: to extract semantic features of Arabic text, propose syntactic patterns of relationships among concepts, and propose a formal model of extracting ontological relations.

The proposed methodology has been designed to analyze Arabic text using lexical semantic patterns of the Arabic language according to a set of features. Next, the features have been abstracted and enriched with formal descriptions for the purpose of generalizing the resulted rules. The rules, then, have formulated a classifier that accepts Arabic text, analyzes it, and then displays related concepts labeled with its designated relationship. Moreover, to resolve the ambiguity of homonyms, a set of machine translation, text mining, and part of speech tagging algorithms have been reused. We performed extensive experiments to measure the effectiveness of our proposed tools. The results indicate that our proposed methodology is promising for automating the process of extracting ontological relations.

1. Introduction

The term Ontology has been defined by Gruber (1993) as "a specification of conceptualization", a formal modeling of a linguistic component along its semantic relationships with respect to other concepts. Ontology can be seen as a pattern of how a given concept is designed to be correlated with other existing ones in a given context.

Developing ontologies from Arabic text is a complex process, as the extraction of the semantic relationships among linguistic components still depends on the syntactic structure of the language. However, interpreting domain independent text requires determining what type of information will be processed and the way it will be expressed. Rather than explaining everything in the text (i.e., syntactic analysis), one could only search for well-known lexical relationships. Thus, meaningful information could be found with simple and soft algorithms, which leads to soft automation of the process.

[*] Corresponding author.
E-mail addresses: Mohammedz@yu.edu.jo (M.G.H. Al Zamil), qasemr@yu.edu.jo (Q. Al-Radaideh).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.007

Table 1 Examples of Hearst-style hyponyms.

Text	Lexical Relation (Hyponym–Hypernym)	Relations
"*Input–output devices, such as Printers*"	Hyponym (Input-Devices, Printer)	NP such as NP
"*Temples are civic buildings*"	Is-a (temple, civic building)	NP is a Adj-Phrase
"*Most European countries, especially France, England, and Spain*"	Kind-of (France, European Country) Kind-of (France, European Country) Kind-of (France, European Country)	NP especially NP

Consider the Hearst-style examples in Table 1 in which a set of useful relations have been extracted in a simple manner. Hearst, (1992) has proposed an acquisition algorithm to detect hyponyms automatically by constructing lexical patterns of knowledge. For instance, consider the example "*input–output devices, such as Printers*" in which a human can easily conclude that printers are a type of input–output device. To make it machine-interpretable, Hearst has proposed the following lexical pattern that can be reused to extract such a relation:

$$NP \text{ such as } \{NP,\}^{*}\{(\text{or}|\text{and})\}NP \tag{1}$$

The disadvantage of Hearst's algorithm is the requirement of collecting real examples as a training set, which is considered a pure supervised activity that requires human intervention. In fact, many researchers found it to be a good feature in that it allows for application of the algorithm to independent text, accents, and special languages by feeding the algorithm with a training set of manual examples. Research on sentimental text analysis in social networks benefits from application of this algorithm, which could be an interesting future direction for Arabic text understanding. Furthermore, specialists in domains such as crime and terrorism detection found this methodology interesting for investigating suspicious text in social networks to detect specific conversations (Ressler, 2006; Salton et al., 1990).

To the best of our knowledge, automatic extraction of semantic relationships from Arabic text based on soft-computing principles has not received significant concern compared with English text. In fact, previous research has focused on the syntactic analysis of Arabic statements and dictionary-based analysis to understand Arabic text for different applications, such as summarization, retrieval, and stemming. Therefore, developing soft and intelligent algorithms for extracting lexical semantic relationships dedicated to Arabic text is of great interest.

In this paper, we consider the application of an enhanced version of Hearst's Algorithm to an Arabic corpus. The proposed methodology has been designed to adapt Hearst's algorithm with additional enhancements to fit our needs, analyzing Arabic text to extract ontological relationships. Such enhancements include: pattern enrichment, pattern filtering, the application of negative patterns, and pattern evaluation. However, experiments have been designed for different types of Arabic text. The results indicated that our proposed methodology is a good candidate to formulate Arabic ontological relations.

This paper is organized as follows: Section 2 discusses the related work in addition to background information about Hearst's algorithm. Section 3 illustrates the framework for applying Hearst's algorithm. Section 4 describes the experiments performed to evaluate the proposed methodology and reports the results. Section 5 discusses and justifies the analysis results. Finally, Section 6 summarizes the conclusions and discusses some future directions.

2. Related work

The most recent decade has witnessed an increasing concern for building Arabic Ontology. Efforts have focused on adapting Arabic ontologies in different natural language processing tasks, such as information retrieval (Moawad et al., 2010), text summarization (Imam et al., 2013), text annotation (Hazman et al., 2012; Dukes and Habash, 2010), improving question answering systems (Abouenour et al., 2008), and building semantic mining of knowledge (Beseiso et al., 2010). The expressive power of the Arabic language makes it difficult to extract ontological relations automatically. Therefore, efficient automatic elicitation of such relations is a complex task that is still reliant on dictionaries (Jarrar, 2013) and cross-language translation, such as Arabic WordNet (Ruiz-Casado et al., 2007; Black et al., 2006; Diab, 2004; Elkateb et al., 2006).

Automatic extraction of ontological relations among language concepts has attracted many researchers. For instance, the ARTEQUAKT project (Alani et al., 2003) has constructed a tool to extract relations to create a biography of a given artist using lexical analysis. Furthermore, many promising techniques have been proposed to handle the problem of creating, managing, and populating Arabic ontology. Al-Yahya et al. (2011) introduced an efficient linguistic approach that restricts its application on fully structured text, such as the Holy Qur'an. Similarly, Al-Rajebah and Al-Khalifa (2013) have incorporated a semantic field in which the meaning of a concept is given according to the concepts around it.

Ghneim et al. (2009) proposed a multilingual framework for Arabic Ontology learning based on previous domain knowledge. A Probabilistic Ontology Model (POM) is applied to represent the extracted ontology. Similar to our proposed technique, the framework learns new concepts and relations using Lexico-syntactical patterns. Another interesting technique is the one proposed by Al-Safadi et al. (2011), which is based on structuring Arabic text into classes, properties, and relationships. The experiments only showed how the developed ontology can be used for querying blogs using Arabic terms. The authors did not provide any experiment regarding the effect of using the proposed ontology for retrieval. While these interesting techniques have introduced ontological relation extraction, we argue that additional enhancements could improve such task.

Practically, there are three methods that have been proposed to automatically extract ontological relations (Wandmacher et al., 2007): repeated-segment, Co-occurrence techniques, and lexical patterns.

The repeated-Segment technique has been applied in Wandmacher et al. (2007), Hernandez (2005). The authors