Journal of King Saud University – Computer and Information Sciences (2014) **26**, 425–440

King Saud University

Journal of King Saud University – Computer and Information Sciences

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

A hybrid method for extracting relations between Arabic named entities

Ines Boujelben *, **Salma Jamoussi, Abdelmajid Ben Hamadou**

Miracl Laboratory, University of Sfax, Tunisia

Available online 28 September 2014

KEYWORDS

Hybrid method;
Relation extraction;
Named entity;
Machine learning;
Genetic algorithm;
Rule-based method

Abstract Relation extraction is a very useful task for several natural language processing applications, such as automatic summarization and question answering. In this paper, we present our hybrid approach to extracting relations between Arabic named entities. Given that Arabic is a rich morphological language, we build a linguistic and learning model to predict the positions of words that express a semantic relation within a clause. The main idea is to employ linguistic modules to ameliorate the results that are obtained from a machine learning-based method.

Our method achieves encouraging performance. The empirical results indicate that the hybrid approach outperformed both the rule-based system (by 12%) and the machine learning-based approaches (by 9%) in terms of the *F*-score, to achieve 75.2% when applied to the same standard testing dataset, ANERCorp.

1. Introduction

Given the enormous amount of Arabic electronic text, we note that there is a high frequency of named entities (NEs) that do not have any linked information. The recognition of these entities represents the first task toward building a semantic analysis and information extraction system. The second task consists of extracting semantic relations between the entities that are useful for a better understanding of human language. Thus, the second task constitutes a crucial move toward natural language processing (NLP) applications. This type of information enables the task of discovering a useful relationship or interaction between two entities from the content of the text. This approach has received a large amount of attention because it is used in many NLP applications, such as automatic summarization, web mining and question–answering (QA).In fact, the NEs' relations extraction can be exploited to extract more precise and correct answers. If we take the example "Where was Taha Hussein born?", the expected answer would be "**Taha Hussein** was born in **Al-Minya Governorate**". The relational triple is born-in (Person, Location), where Person and Location are the NEs.

Therefore, several studies on NE recognition have already been performed in many languages, such as English, French and Chinese. Additionally, many NE recognition systems have been built for the Arabic language. In the literature, three types of approaches have been proposed for Arabic NE recognition systems. Some of the proposed systems rely on handcrafted

* Corresponding author.
E-mail addresses: Boujelben_ines@yahoo.fr (I. Boujelben), jamoussi@gmail.com (S. Jamoussi), adelmajid.benhamadou@isimsf.rnu.tn (A. Ben Hamadou).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.004

rules, namely, the rule-based approach (Mesfar, 2007) and (Fehri et al., 2011). Other studies use a machine learning (ML)-based approach. They utilize a set of features that were extracted from an annotated corpus. In this context, (Benajiba and Rosso, 2008) and (Abdul-Hamid and Darwidh, 2010) have used Conditional Random Fields sequence labeling. (Benajiba and Rosso, 2008) reported 90%, 66% and 73% *F*-measures for the location, organization and persons, respectively. (Abdul-Hamid and Darwidh, 2010) achieved an improvement in the *F*-measure over (Benajiba and Rosso, 2008) for recognizing persons and organizations, by 9 points and 2 points, respectively. Finally, a few studies in Arabic NE recognition have used a mixed approach. We mention (Shaalan and Oudah, 2014), who concentrated on a hybrid approach. Because of their combination of rule-based and ML-based approaches, these authors achieved a 90% *F*-measure. Their system outperforms the state-of-the-art for Arabic NER in terms of accuracy when applied to the ANERCorp[1] standard dataset.

However, the results reported in the NE relation extraction task were not as good as those achieved in the NE recognition task. For this task, only a few studies have addressed the Arabic language. We notice (Ben Hamadou et al., 2010a), whose approach is based on patterns that were rewritten into local grammar within the linguistic platform NooJ. They aimed to extract functional relations between persons and organizations. Additionally, (Alotayq, 2013) adopted the learning classifier MaxEnt to extract relations between various types of NEs. To the best of our knowledge, there is no study that has adopted a hybrid approach to discover the relations between NEs in the Arabic language. Thus, it will be challenging to adopt this approach for extracting the relations between NEs in the Arabic language.

In this paper, the relations between Arabic NEs are tackled through developing a hybrid system to combine the advantages of ML- and rule-based approaches. Mainly, an ML approach followed by a post-processing rule-based approach is used in an attempt to enhance the overall performance of the ML system. Our aim is to predict the trigger words that express the semantic relations between NEs from Arabic text, relying on a set of rules. First, our system is based on ML algorithms to extract the rules using a decision tree technique and an Apriori algorithm. Then, a genetic algorithm (GA) is used to extract and generate the most significant and interesting rules. After applying ML methods, we added hand-crafted rules to treat both invalid examples and unseen relations.

The remainder of this paper is organized as follows: First, we survey prior studies on relations extraction. Section 2 provides background on relations between NEs. Then, we explain the relation extraction task as well as the different challenges. The fourth section illustrates the architecture of our hybrid process, in which we detail the main steps of our proposed method. Afterward, we present the different experiments from which we discuss the reported results.

2. Related studies

Today, relation extraction that involves NEs is seen as a step toward a more structured model of text meaning. Several methods have been proposed to extract semantic relations between NEs. These methods can essentially be classified into three broad categories: the rule-based approach, ML-based approach and hybrid approach.

2.1. Rule-based approach

In the first approach, the rules are usually implemented in the form of regular expressions or finite-state transducers. From the studies performed in the Arabic language, we mention (Ben Hamadou et al., 2010a) and (Boujelben et al., 2012).These authors extracted a set of linguistic patterns from a training corpus. Subsequently, they rewrote those patterns into finite state transducers within the linguistic platform NooJ,[2] using specifically local grammars.[3] This approach uses a representation of linguistic rules by means of transducers.

(Ben Hamadou et al., 2010a) reported an *F*-score of 70%, while (Boujelben et al., 2012) achieved an *F*-score of 60%. This result is significant because (Ben Hamadou et al., 2010a) is limited to only the functional relations between the NE pairs (PERS–ORG). Thus, they concentrate solely on one NE pair, which enables them to construct more precise and concise rules. In contrast, (Boujelben et al., 2012) are interested in extracting more relations among five pairs of NEs (PERS–LOC, PERS–PERS, PERS–ORG, ORG–LOC and LOC–LOC). To extract the relations between these NE pairs, the authors elaborated five sub-grammars. Each grammar contains the pattern of relations between each pair. The system considers the gender and the number features of the relation triggers when it verifies whether the NEs are related. Because of these NooJ grammars, their process enables the extraction of semantic relations that are predicted through one or multiple word forms that appear before, between, or after the NEs.

The rule-based method offers a significant analysis of the context for each NE and its relations with the other NEs. However, the complexity of Arabic sentences and the high variability in the expressions used make it intricate to detect some of the relations between the NEs. To accomplish that goal, a tangible effort is required to write down all the rules for discovering relations between NEs. To overcome this manual step, some studies, such as (Ezzat, 2010), are oriented to a semi-automatic method for automatically producing recognition grammars for relation detection between NEs. These grammars present a set of patterns that are provided by an algorithm. The algorithm relies on generalizing a large collection of sentences that contain the relevant relation. These sentences are collected by a linguist or a domain expert.

2.2. Machine learning-based approach

To fully automate the relation extraction task, some research studies have been oriented toward ML methods, including un-supervised, semi-supervised and supervised learning techniques.

The un-supervised methods make use of massive quantities of unlabeled text and are based almost entirely on clustering

[1] Available on http://www1.ccls.columbia.edu/~ybenajiba/downloads.html.

[2] Available on http://www.nooj4nlp.net.

[3] NooJ local grammars are typically used to describe sequences of words that present meaningful units or entities. In fact, these grammars can be used to locate syntactic constructions of interest, such as sentences that contain certain grammatical words or syntactic constructs.