Journal of King Saud University – Computer and Information Sciences 29 (2017) 212–228

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

# Deriving ontological semantic relations between Arabic compound nouns concepts

Imen Bouaziz Mezghanni *, Faiez Gargouri

*MIR@CL Laboratory, Sfax University, Tunisia*

ARTICLE INFO

*Article history:*
Received 6 November 2016
Revised 27 February 2017
Accepted 1 March 2017
Available online 10 March 2017

*Keywords:*
Semantic relations derivation
Arabic compound nouns
Compound structure
Gamification
FCA
RCA

ABSTRACT

Legal ontologies have proved their increasingly substantial role in representing, processing and retrieving legal information. By using the knowledge modeled by such ontologies in form of concepts and relations, it is possible to reason over the semantic content of legal documents. Supporting (semi-) automatically the development of ontologies from text is commonly referred to as ontology learning from text. The learning process includes learning of the concepts that will form the ontology and learning of the semantic relations among them.

In this paper, we present a new approach for expliciting the semantic relations between Arabic compound nouns concepts. The originality of this work is twofold. Firstly, the technique of inferring relations is based on exploiting the internal structure of the compounds using a defined set of domain-and language-independent rules according to their different structures, on the one hand, and on studying prepositions semantics specifying the inferred relations applying a gamification mechanism that collects human votes, on the other hand. Secondly, relying on the compounds set described by both binary (structural positions in which there are written) and relational attributes (the deduced relations), we used a "Relational Concept Analysis" (RCA) technique, as an adaptation of "Formal Concept Analysis" (FCA), for the construction of interconnected lattices that we transformed into ontological concepts and relations which can be either taxonomic or transversal.

Experiments carried out on Arabic legal dataset showed that the proposed approach reached encouraging performance through achieving high precision and recall scores. This performance affects positively the retrieval results of legal documents based on a powerful ontology, which presents our main objective.


## 1. Introduction

With the development of information technology and easier internet access, electronic dissemination of huge amounts of published documents has made the legal information retrieval more and more complex for the user. Nowadays, search engines present the main tools for accessing data available on the Web. However, most search engines do their text query and retrieval using keywords, which often results in hits completely irrelevant to user query leading to low precision and recall parameters. The weakness of search engines can be overcome through using Semantic Web technologies considered to be the next generation of the actual Web. The Semantic Web is a Web of ontologies that allow the analysis of the domain knowledge by modeling the relevant concepts of this domain. The ontologies enable semantic interoperability involving the comprehension of information to be precisely described and well understood by machine. Therefore, the search is no longer based on keywords matching, but rather on concepts matching. In this case, the search results become more relevant, which increases precision and recall rates.

However, the manual building of ontologies is a time consuming and labor intensive task. Ontology learning (Maedche and Staab, 2004), which aims at providing automatic and semi-automatic approaches for ontology generation, can overcome the bottleneck of knowledge acquisition. The learning process includes learning of the concepts that form the ontology and learning of the semantic relations among them. This paper introduces a novel approach for expliciting the semantic relations between Arabic compound nouns concepts.

To further explain the proposed approach, it is necessary to define the following terms:

* Corresponding author.
*E-mail address:* imen_bouaziz_miracl@yahoo.com (I.B. Mezghanni).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.03.001
1319-1578/

- "A term is a lexical unit consisting of one or more than one word which represents a concept inside a domain" (de Bessé et al., 1997).
- "A concept is an abstract unit which consists of the characteristics of a number of concrete or abstract objects which are selected according to specific scientific or conventional criteria appropriate for a domain" (de Bessé et al., 1997).
- "A multi-word term or compound term is a combination of a set of words used to convey a single unit of meaning. Its semantics depends on the knowledge area of the concept it describes and cannot be inferred directly from the semantic composition of its components separately" (Sag et al., 2002).

The association "term = concept" is erroneous. Indeed, a term can represent many concepts. For example, the term "draft" can refer either to a current of air into an enclosed space or to the first version of a document, plan or drawing. However, a concept may be denoted by many terms. Therefore, terms are considered as units of language, while concepts are elements of the conceptual model.

Whatever the text and the language in which the compounds are written, they are often viewed as relevant since they play an important role in the encapsulation and expression of nominal concepts. Compounds are also frequent in a wide variety of texts types, which makes their extraction a crucial task.

In a conceptual model, considering compound concepts without taking into account a predefined relation linking them is not very significant as it may lead to their discard. Determining the semantic relations between concepts is fundamental in capturing the ideas in texts. Besides, relations, such as part-whole, cause-effect, etc., encode crucial information about how different entities should be perceived in relation to each other. Thus, much attention has been paid to this research field and several works have recently been carried out on different languages, such as English (Ta and Thi, 2016; Joseph et al., 2016) and Chinese (Miao et al., 2012), etc.

For instance, extracting semantics from compound nouns was tackled by Vela and Declerck (2009) in a process of ontology building. Relying on pattern-based approaches, compounds were first detected and analyzed to suggest candidate ontology classes and relations. Then, paraphrases of the compounds in the text were detected through a set of patterns and analyzed in order to filter and validate the list of candidate classes and relations obtained in the first step. In their approach, only noun-noun compounds were taken into account.

With the intention of automatic Thai ontology construction, Kawtrakul et al. (2004) processed parses sentences and generated compound nouns as candidate terms based on phrase chunking. Using statistical-based technique, the compounds were analyzed in order to separate head and modifier from each other. The semantic relations of a compound were extracted by learning the common ancestral concept of its head and modifier using heuristic rules as well as expert's judgments.

Sruti Rallapalli (2012) explored the scope of identifying the semantic relation. Thereby, he interpreted compound nouns using an indexed semantic ontology combined with noun similarity measurement techniques. The problem, here, is that the semantic similarity is limited to the ontology itself as the primary information source. Therefore, there is a need for creating standard corpora in any domain of application.

Extracting semantic relations from compound nouns can be also based on a frame-semantic approach (Lakhfif and Laskri, 2016). The basic idea of the latter is that meanings, such as purpose, constitution and agency, their realization,etc., can be viewed as a generalized and lexicalized aspect of qualia structure as defined by Pustejovsky (1991). In this context, the challenge consists in the ability to organize relational possibilities hierarchically according to the compounds underlying semantic meanings and the ability to recognize an implication structure among different but related relational possibilities.

However, despite the importance of the Arabic language, few studies investigated the process of extracting the semantic relations in Arabic texts due to the complexity of this task. This complexity arises from the distinctive features characterizing the Arabic language, namely the agglutination and diactritization causing major morphological and syntactic ambiguities.

In this paper, the derivation of semantic relations between Arabic compound nouns is dealt with through developing a hybrid approach to combine the advantages of clustering and rule-based approaches. From the compounds, two kinds of implicit relations (is-a relation; objectProperty relation) are extracted based on a set of pattern rules defined according to the different structures of the compounds. To specify the resulting objectProperty relation, we resorted to prepositions in order to describe the hidden relations present in the compounds through a gamification mechanism. Gamification refers to "the use of design elements characteristic for games in non-game contexts" (Deterding et al., 2011). A validation step was followed by experts in order to verify the accuracy of the chosen relations and the reliability of the proposed rules.

This work introduces a part of ontology construction whose goal is to support retrieval of legal documents and in which we focus on the structural positions where the concepts appear in the document. This position is determined by referring to such structural element of the document. In a legal code, we considered a structural position, the article number to which a concept belongs (Mezghanni and Gargouri, 2015). Thus, a concept is described by the "Articles" where it is written. For instance, the concept حاكم التحقيق (investigating judge) is described by article 10 and article 11. The ontological concepts together with their associated positions are defined by means of an incidence matrix to FCA (Ganter and Wille, 1997; Ganter et al., 2005) which is a mathematical approach for data analysis providing a rigorous framework for the derivation of a conceptual hierarchy called "concept lattice".

In order to handle the generated relations between concepts (the concept حاكم التحقيق (investigating judge) is-a حاكم (judge)), we relied on RCA (Huchard et al., 2007, 2003) as an extension of FCA which includes further relational structures. Indeed, RCA considers the relations between objects in addition to the characteristics of the objects (sets of object-attribute data provided with relations). In other words, objects are described by attributes and their relations with other objects. RCA consists in iteratively applying an FCA algorithm using relational data. The discovered concepts at a given step are propagated along the relations, leading to the discovery of new concepts at the next iteration.

The remainder of the paper is organized as follows: Section 2 discusses recent works in the domain of semantic relation extraction from Arabic texts. Then, we recall the basic notions of ontologies, FCA and RCA in Section 3. The adopted approach is described in Section 4. Section 5 shows the experiments and the obtained results evaluated in Section 6. A conclusion with future research directions are presented at the end of the paper.

## 2. Related work

In the literature, several researches were conducted to investigate the process of Arabic ontologies learning in different applications. These ontologies belong to various domains and were constructed differently. The survey proposed in Mezghanni and Gargouri (2015) summarizes recent works presented for ontology learning from Arabic textual resources.