Engineering Applications of Artificial Intelligence 51 (2016) 230–241

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Generating core domain ontologies from normalized dictionaries

Feten Baccar Ben Amar [a], Bilel Gargouri [a,*], Abdelmajid Ben Hamadou [b]

[a] MIRACL Laboratory, FSEGS, B.P. 1088, 3018, University of Sfax, Tunisia
[b] MIRACL Laboratory, ISIMS, B.P. 242, 3021 Sakiet-Ezzit , University of Sfax, Tunisia

ARTICLE INFO

Available online 2 March 2016

Keywords:
Domain ontology
Core generation
LMF (ISO 24613) standardized dictionary
Ontology consistency
General framework

ABSTRACT

This paper proposes a general framework for automatic core domain ontology generation from LMF (ISO 24613) standardized dictionaries. The originality of this work lies not only in the use of a unique and finely structured source containing multi-domain and lexical knowledge of morphological, syntactic and semantic levels, lending itself to ontological interpretations, but also in the proper building of the taxonomic backbone of the domain ontology. To this end, we have integrated a validation stage into the proposed process in order to maintain the consistency of the resulting formalized domain ontology core throughout this process and support the checking of anomalies in the handled source. Furthermore, this generation process has been implemented in an iterative and incremental system based on domain- and language-independent rules. The reliability of the proposed process is proven through many experiments that have been conducted on various domains using normalized dictionaries, but without lack of generality, we choose to illustrate an experiment carried out on the Arabic language. This choice is explained by both the great deficiency of work on building of Arabic ontologies and the availability within our research team of an LMF-standardized Arabic dictionary.

1. Introduction

In recent years, research on ontology development and construction process improvement has become increasingly widespread in computer science community. Indeed, domain ontologies are extremely powerful knowledge representation tools for describing a set of relevant domain-specific concepts and their relationships in a formal way (Guarino, 1998). Although the field of ontology learning aiming to automate the ontology creation process has been dealt with by plenty of work, it is still a long way from being fully automatic and deployable on a large scale. Actually, it requires significant human (expert) involvement for the validation of each step throughout this process (Lonsdale et al., 2010).

In order to reduce the costs, research on ontology learning has been conducted using a variety of resources, such as raw text (Poon and Domingos, 2010 ; Aussenac-Gilles et al., 2008; Li et al., 2005; Navigli et al., 2003), XML structured data (Aussenac-Gilles and Kamel, 2009, Bedini et al., 2011), Machine-Readable Dictionaries (MRDs) (Kurematsu et al., 2004; Kietz et al., 2000; Rigau et al., 1998), and thesauri (Li and Li, 2012; Chrisment et al., 2008; Soergel et al., 2004). Obviously, these resources have different features, and therefore, each proposed process is based on a different approach pertaining to rules, Natural Language Processing (NLP) techniques, etc.

As linguistic information is increasingly required in ontologies mainly in NLP applications (Buitelaar et al., 2009; Pazienza et al., 2007), among the considered terminological resources, MRDs represent one of the most likely and suitable sources promoting the knowledge extraction both at conceptual and lexical levels. However, since much lexical information has not yet been encoded, access to the potential wealth of information in dictionaries remains limited for software applications.

From another standpoint, the growing awareness of the benefits of having finely structured knowledge in lexical resources has led to the definition of Lexical Markup Framework (LMF) (ISO 24613, 2008). Its meta-model basically provides a common and shared representation of lexical objects that allows the encoding of rich linguistic information, including morphological, syntactic and semantic aspects (Francopoulo and George, 2008). Particularly, an LMF-standardized dictionary incorporates widely-accepted and commonly-referenced diversified linguistic knowledge lending itself to ontological interpretations. Besides, finely structured and multi-domain knowledge in an LMF-standardized dictionary paves the way for automatically generating ontological entities that

* Corresponding author.
E-mail addresses: feten.baccar@mes.rnu.tn (F.B.B. Amar), bilel.gargouri@fsegs.rnu.tn (B. Gargouri), abdelmajid.benhamadou@isimsf.rnu.tn (A.B. Hamadou).

http://dx.doi.org/10.1016/j.engappai.2016.01.014
0952-1976/

constitute the core of the targeted domain ontology (Baccar et al., 2010).

The ultimate objective of this paper is to propose a framework for core domain ontologies generation starting from LMF-standardized dictionaries. In fact, the systematic organization of such resource allowed us to implement a fully automatic process for a direct dictionary transformation of some particular information into ontological elements based on domain- and language-independent rules. In addition, since evaluating ontology as a whole is a costly and challenging task, especially when the reduction of human intervention is sought (Almeida, 2009) a validation stage had to be integrated into our iterative process. Indeed, the detected errors would be avoided in ontology to maintain its consistency, on the one hand, and should be reported back to an expert to indicate anomalies in the handled source, on the other hand. Then, the interest of this stage is twofold: first, it guarantees the quality of the produced ontology; second, it contributes to the checking of anomalies in the handled dictionary that are hard to be detected manually (Wali et al., 2014).

Apart from its generality and fully automated level, the proposed framework has the merit to support under-resourced languages, such as Arabic language, in the sense that new resources could be generated from existing ones with the least effort and costs.

Furthermore, the implemented system has proven to be trustworthy through a series of experiments that have been conducted on various domains using normalized dictionaries, but without lack of generality, the experiment reported in this paper was carried out on the Arabic language. This choice is explained not only by the great deficiency of work on Arabic ontology building, but also by the availability within our research team of an LMF-standardized Arabic dictionary (Baccar et al., 2008).

Concerning the ontology implementation, we chose to use the Ontology Web Language (OWL) language in its version 2 (Motik et al., 2009). It is a formal and standard language proposed to represent ontologies for the Semantic Web. Indeed, the use of this language is stimulated by its reasoning capabilities and standard nature.

The remainder of this paper is structured as follows. Section 2 presents the state-of-the art and the motivations of our work. Section 3 describes the proposed framework for the core domain ontology generation from normalized dictionaries. Then, Section 4 gives details of the system implementation and its experimentation. Next, in Section 5, we discuss the assessment of the proposed framework as well as the quality of the obtained experimental results. Finally, Section 6 concludes the paper and opens perspectives for future work.

2. State-of-the art and motivations

Ontology engineering has always been a tedious task requiring considerable human involvement and effort especially in the activity of knowledge acquisition. During the three last decades, there have been some efforts to automate ontology construction process by exploiting structured documents like XML structured data sources (Aussenac-Gilles and Kamel, 2009). However, almost all the proposed approaches suffer from many drawbacks, especially their non generality (narrow scope of application) and their reliance on human intervention (Bedini et al., 2011).

Another interesting kind of (semi-) structured knowledge content to learn ontology is the one provided by Machine-Readable Dictionaries (MRDs). In fact, early work on extracting taxonomies from MRDs goes back to the 80s and early 90s (Michiels et al., 1980; Amsler, 1981; Calzolari, 1984; Chodorow et al., 1985; Dolan et al., 1993; Wilks et al., 1996). Bearing in mind that MRDs can be exploited to obtain rich and explicit semantic information between lexical entries, these researchers have attempted to make explicit the implicitly embedded knowledge in the definition texts. The basic idea is to exploit the regularity of dictionary entries to initially find a suitable *hypernym* for the defined word (For more details, we refer the reader to (Cimiano, 2006)).

Furthermore, considered as a large repository of semi-structured knowledge about a language and about information related to the real world, MRDs have been the backbone for generating conceptual structures ranging from concept hierarchies (Jannink and Wiederhold, 1999; Yamaguchi, 1999), thesauri (Jannink, 1999) to ontologies (Kurematsu et al., 2004; Nichols et al., 2005). Indeed, the main advantage of using existing human-oriented knowledge resources such as MRDs, is the possibility to exploit their partial structure. Actually, word senses, which are separately defined in MRDs, can be seen as the equivalent of ontological categories, and semantic relations (e.g., *synonymy*, *antonymy*, *hyponymy*, *meronymy*) between the different senses would correspond to ontological relations (for example, *hypernym/hyponym relation* would stand for *subsumption*) (Hirst, 2004). Another important aspect is that ontological relationships are learned between word senses (i.e. concepts) rather than between the words themselves (Cimiano, 2006).

Nevertheless, since the MRDs are oriented towards the human reader, much information is not well-structured and subsequently its machine interpretation might not be evident. Consequently, systems relying on MRDs encounter two major problems. While the first inherent problem is the need for massive human intervention, the second is their confinement to limited relations in almost all but the taxonomic ones (Vaquero et al., 2007).

A common and unified model for the creation and use of computational lexicons, baptized Lexical Markup Framework (LMF), has been defined (ISO 24613, 2008). It should be noted that even though its main goals are to manage the exchange of data among lexical resources and enable their merging at small or large scale (Francopoulo and Georges, 2008), the LMF meta-model allows for transforming MRDs' content into a finely structured and more explicit knowledge format, thus facilitating access to its manifold lexical information. Indeed, covering all the natural languages (including languages with rich and complex morphology, such as Arabic), the LMF meta-model contains much explicit linguistic information (inflectional and variant forms, synonyms, part-of-speech, definitions, usage domain, grammatical properties, etc.) as well as a lot of semantic knowledge disseminated in the definitions and examples.

From another standpoint, the growing awareness of the benefits of the linguistically grounded ontologies, especially in semantic web and NLP communities, has led to the creation of a number of models attempting to extend ontological objects with linguistic grounding (Pazienza et al., 2008; Aussenac-Gilles et al., 2008). Moreover, after the introduction of LMF standard, a good deal of active work, among which we can mention LexInfo (Buitelaar et al., 2009) and Linguistic Information Repository (LIR) (Montiel-Ponsoda et al., 2008), has been undertaken in response to the need for increasing the linguistic expressivity of given ontologies. The proposed models try to associate lexical information with ontological entities in the same resource the making of which would be a heavy and time-consuming activity due to the plurality and the heterogeneity of the handled sources. In addition, some complexity arises when linguistic information is involved in ontology reasoning (Ma et al., 2010).

Since an undeniable contribution of the LMF standard lies in its ability of rich representation of MRD and NLP lexica, which enables a direct and selective access to various lexical objects in the linguistic resources, we have asserted that an LMF-standardized