Expert Systems with Applications 40 (2013) 2058–2068

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A semantic role labelling-based framework for learning ontologies from Spanish documents

José Luis Ochoa [a,*], Rafael Valencia-García [b,1], Alonso Perez-Soltero [a], Mario Barceló-Valenzuela [a]

[a] Department of Industrial Engineering, Universidad de Sonora, Blvd. Rosales y Transversal, Hermosillo, Sonora, C.P. 83000, Mexico
[b] Faculty of Computer Science, Universidad de Murcia, 30071 Espinardo (Murcia), Spain

ARTICLE INFO

Keywords:
Ontology learning
Semantic role labeling
Information extraction
Semantic Web
Ontology

ABSTRACT

Currently, most of the information available in the Web is adapted primarily for human consumption, but there is so much information that can no longer be processed by a person in a reasonable time, either in digital or physical formats. To solve this problem, the idea of the Semantic Web arose. The Semantic Web deals with adding machine-readable information to Web pages. Ontologies represent a very important element of this web, as they provide a valid and robust structure to represent knowledge based on concepts, relations, axioms, etc. The need for overcoming the bottleneck provoked by the manual construction of ontologies has generated several studies and research on obtaining semiautomatic methods to learn ontologies. In this sense, this paper proposes a new ontology learning methodology based on semantic role labeling from digital Spanish documents. The method makes it possible to represent multiple semantic relations specially taxonomic and partonomic ones in the standardized OWL 2.0. A set of experiments has been performed with the approach implemented in educational domain that show promising results.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Traditionally, the information stored in the Web has been created and designed for human consumption (Berners-Lee, 1998), however, in recent years the trend has changed and the information of the Web must also be processed automatically by computers. Since, although consumption is still human, information recovery, extraction and processing should be done by machines to liberate ourselves from the shackles that represents acquiring valid information. Thus, manual work is required for making that information machine-readable, which can be tedious, difficult, and time-consuming (Han & Elmasri, 2004).

In Berners-Lee, Hendler, and Lassila (2001) the Semantic Web was defined as an extension of the current Web in which information is provided with well-defined meaning, so that computers and people can work in a cooperative manner. In the Semantic Web, ontologies are used as knowledge representation technology. In this work, an ontology is viewed as a formal specification of a shared domain knowledge conceptualization (Studer, Benjamins, & Fensel, 1998). In this sense, ontologies provide a formal, structured knowledge representation, having the advantage of being reusable and shareable. In our methodology, ontologies are used to represent the knowledge extracted from texts, so that ontologies are obtained as a result of knowledge extraction processes. More concretely, the second version of the Ontology Web Language OWL 2 is used (Grau et al., 2008). OWL 2.0 provides several new features to OWL, including a high expressive power for properties and extended support for datatypes.

Ontologies are currently being applied to a number of different domains such as bioinformatics (Chen, Huang, Bau, & Chen, 2012), medicine (Arsene, Dumitrache, & Mihu, 2011), tourism (Jung, 2011), software engineering (Boskovic et al., 2011) and Cloud Computing (Jimenez-Domingo, Gomez-Berbis, Colomo-Palacios, & Garcia-Crespo, 2011).

Due to the outstanding importance of ontologies in these areas and in the Semantic Web, different methodologies for building and developing ontologies from scratch have been proposed in the last years (Ruiz-Martínez, Valencia-García, Fernández-Breis, García-Sánchez, & Martínez-Béjar, 2011). On the other hand, the manual construction of such ontologies is a major problem, since it consumes both time and resources (Ruiz-Martínez et al., 2011), so other mechanisms such as ontology learning, are necessary to support the construction of ontologies. Ontology Learning is a knowledge acquisition activity used to transform data sources into ontologies. The majority of the approaches of ontology learning deal with the ontology construction from natural language text and more concretely the vast majority of these ontology learning methods have focused on the English language. Nevertheless, the

[*] Corresponding author. Tel.: +52 (662) 259 21 59; fax: +52 (662) 259 21 60.
E-mail addresses: joseluis.ochoa@industrial.uson.mx (J.L. Ochoa), valencia@um.es (R. Valencia-García), aperez@industrial.uson.mx (A. Perez-Soltero), mbarcelo@industrial.uson.mx (M. Barceló-Valenzuela).
[1] Tel.: +34 868888522; fax: +34 868884151.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.017

Spanish language has a much more complex syntax, and is nowadays the third most spoken language in the world, for which we firmly believe that the computerization of Internet domains in this language is of highest importance.

In this paper, we propose an automatic and domain independent method for ontology learning from Spanish natural language texts based on the identification of semantic relations among concepts using semantic roles.

The rest of the paper is organized as follows. The related work is described in Section 2. The proposed method is explained in Section 3. A validation of the ontology learning method in an universitary domain corpus is described in Section 4. Finally, conclusions and future work are defined forward in Section 5.

2. Related work

The development of the Semantic Web together with the growth of textual and ontological resources in the Web have taken that the efficient automatic building and maintenance of the knowledge repositories, such as ontologies, be a burning issue (Spasic, Ananiadou, McNaught, & Kumar, 2005). Although in the last decade several systems for learning ontologies have been proposed, most of them are focus on English language and have important drawbacks, since they are only capable of extracting taxonomies or significantly reduced sets of relations and their degree of automation (Maedche & Staab, 2001). Actually, the linguistic features of texts from which the ontologies are extracted make difficult the development of systems of general purpose. This difficulty increases in specific domains, such as the biomedical domain, that makes use of a specific sublanguage (Ananiadou & Mcnaught, 2006; Friedman, Kra, & Rzhetsky, 2002).

Table 1 shows some of the main ontology learning systems applied to different domains. The classification is an adaptation of the classification presented in Petasis et al. (2007).

The main parameters are Initial requirements, Degree of automation, Domain portability, Relationships supported, Consistency maintenance and Language dependence.

2.1. Initial requirements

Concerning the initial requirements, namely resources or background knowledge, most of the approaches make use of terminology extraction modules. Terminology extraction is a subtask of information extraction whose goal is to automatically extract relevant terms from a given corpus. There are three main approaches to term extraction (Ochoa, Almela, & Valencia-García, 2011b). The linguistic approach relies on the assumption that terms have typical morphological features and recurrent syntactic structures. The statistical approach assumes that terms have statistical features which are different from normal words. The third strand of research, the hybrid approach, relies on the assumption that combining linguistic and statistical approaches in various stages of the process of term extraction can provide more accurate results.

In ontology learning, the statistical approach is the most currently used. Meanwhile the TF-IDF algorithm is the most commonly used (Villaverde, Persson, Godoy, & Amandi, 2009), other approaches use the random walk term weighting (Hou, Ong, Nee, Zhang, & Liu, 2011) or co-occurrence measures (Sánchez, Moreno, & Vasto-Terrientes, 2012). On the other hand, linguistic approaches are used in multitude of approaches such as Navigli and Velardi (2004), Ruiz-Martínez, Valencia-García, Fernández-Breis, García-Sánchez, and Martínez-Béjar (2011), Zouaq, Gasevic, and Hatala (2011) and Dahab, Hesham, and Hassan (2008). Finally, other works such as TextToOnto can be configured to use different statistical and linguistic approaches (Cimiano & Volker, 2005). In

Table 1
Ontology Learning methodologies.

Approach	Initial requirements	Learning approach	Degree of automation	Domain portability	Consistency maintenance	Relationship supported	Language dependency
Villaverde et al. (2009)	Statistical terminology extraction – TF-IDF	Pattern-based	Semi-automatic	Domain specific	No	Non-taxonomic	Yes (English)
Hou et al. (2011)	Statistical terminology extraction – TF-IDF	Pattern-based	Automatic	Portable	No	Non-taxonomic	Yes (English and Chinese) Fairly portable
Dahab et al. (2008)	Linguistic terminology extraction	Pattern-based	Automatic	Portable	No	Non-taxonomic	Yes (English) Portable
Zouaq et al. (2011)	Linguistic terminology extraction	Pattern-based	Semi-automatic	Portable	No	Non-taxonomic	Yes (English)
Ruiz-Martínez et al. (2011)	Linguistic terminology extraction	Rule-based	Semi-automatic	Domain specific	Yes	Taxonomic Partonomic Dependence Topologic Causal Functional	Yes (English) Portable
Sánchez et al. (2012)	Statistical terminology extraction - Co-ocurrence	Pattern-based	Automatic	Portable	No	Non-taxonomic	Yes (English) Portable
Navigli and Velardi (2004)	Linguistic terminology extraction	Inductive machine learning	Automatic	Portable	Yes	Taxonomic Partonomic	Yes (English) Portable
Cimiano and Volker (2005)	Hybrid terminology extraction	Machine learning	Semi-automatic	Portable	No	Taxonomic Partonomic	Yes (English) Portable
Our proposal	Hybrid terminology extraction – NC-Value – TF-IDF	Semantic role labelling Pattern-based	Automatic	Portable	Yes	Taxonomic Partonomic Other semantic relationships	Yes (Spanish) Farily portable

Download English Version:

https://daneshyari.com/en/article/10321959

Download Persian Version:

https://daneshyari.com/article/10321959

Daneshyari.com