Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X
PP: 5-12

Automatic Learning of Ontologies for the Semantic Web: experiment lexical learning[1]

Eduard Puerto[2] | Jose Aguilar[3] | Taniana Rodriguez[4]

Recibido:
Abril 20 de 2012

Aceptado:
Junio 2 de 2012

Abstract

This paper proposes the design of a System for Automatic Learning of Ontologies and Lexical Information (SALOX) for the Dynamic Semantic Ontological Framework for the Semantic Web (DSOFSW). DSOFSW interprets query in natural language (Spanish) to the web, and is composed by five parts; a linguistic ontology for the grammar of Spanish, a lexicon for the lexical information, a database of facts about the system experiences, a task ontology for the linguistic analysis process, and an interpretative ontology of the context. SALOX integrates several methods, approaches and techniques for information extraction, discovery and actualization (pragmatic (user profile, context knowledge), lexical and semantic linguistic information, etc.) in order to update the knowledge used for DSOFSW. SALOX has a component to map the sources of learning with the learning methods, and another to update the linguistic ontology and the lexicon of the DSOFSW. Specifically, in this paper we present the design of the learning unit of lexical information.

Keywords: *Natural language processing, ontological semantic, machine learning, ontological leaning, semantic web.*

Resumen

Este artículo propone el diseño de un sistema para el aprendizaje automático de ontologías e información léxica (System for Automatic Learning of Ontologies and Lexical Information - SALOX) para un Marco Ontológico Dinámico Semántico para la Web Semántica (Dynamic Semantic Ontological Framework for the Semantic Web – DSOFSW). DSOFSW interpreta consultas en lenguaje natural (español) para la Web, y está compuesta por cinco partes: una ontología lingüística para la gramática del español, un lexicón para la información léxica, una base de datos de hechos sobre el sistema de experiencias, una ontología de tareas para los procesos de análisis lingüísticos, y una ontología interpretativa para el contexto. SALOX integra varios métodos, enfoques y técnicas para la extracción de información, descubrimiento y actualización (pragmática (perfil

[1]Acknowledgements. Thanks to the CDCHT Project I-1237-10-02-AA of the Universidad de los Andes, to the 2994 project of FONACIT-PEII, and to the project "Motor de búsqueda basado en ontologías" of the Universidad Francisco de Paula Santander.

[2]Universidad Francisco de Paula Santander, Dpto de Sistemas e Informática. I.S, M.Sc en Ciencias de la computación. Cúcuta, Colombia
e-mail: eduardpuerto@ufps.edu.co

[3]Universidad de Los Andes(República Bolivariana de Venezuela) Facultad de Ingeniería, Sector la Hechicera, Director del CEMISID (Centro de Investigaciones en Microcomputación y Sistemas Distribuidos)
Ph.D en Inteligencia Artificial
e-mail: aguilar@ula.ve,

[4]Universidad de Los Andes (República Bolivariana de Venezuela)Facultad de Ingeniería, I.S, Estudiante de Doctorado en Ingeniería.
e-mail: taniana@ula.ve

Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X

de usuario, conocimiento de contexto), información léxica y de lingüística semántica, etc.) con el fin de actualizar el conocimiento usado para DSOFSW. SALOX tiene un componente que mapea las fuentes de aprendizaje con los métodos de aprendizaje, y otro que actualiza la ontología lingüística y el lexicón del DSOFSW. Específicamente, en este artículo presentamos el diseño de la unidad de aprendizaje de información léxica.

Palabras clave: *Procesamiento de lenguaje natural, semántica ontológica, aprendizaje de máquina, aprendizaje ontológico, web semántica.*

1. Introduction

The Web is evolving from a huge information and communication space into a massive knowledge and services repository. One of enablers of the above change is the ontology, commonly referred to as the conceptualization of a domain. Ontology provides a semantic base for the machine-understandable description of digital content [1]. It is ubiquitous in information systems by annotating documents with meta-dada, improving the performance of information retrieval and reasoning, and making data between different applications interoperable. In addition, ontology-type semantic description of behaviors and services allows software agents in a multi-agent system to better coordinate themselves. Therefore, ontology development will have a profound impact on a wide range of enterprise applications and knowledge integration [2].

Ontology learning refers to the automatic discovery and creation of ontological knowledge using machine learning techniques. Ontology learning uses methods from diverse fields such as machine learning, knowledge acquisition, natural language processing, information retrieval, artificial intelligence, reasoning and database management. Ontology learning seeks to discover ontological knowledge from various forms of data automatically or semi-automatically, using previous methods [3].

In this paper we propose a learning architecture that integrates several of these methods in a same system to support the process of interpretation of a query in natural language (Spanish). Additionally, we show in detail the design of one of the learning units, for the case of lexical information. The article is structured as follows. Section 2 presents the Background, section 3 provides a brief description of DSOFSW, section 4 presents the SALOX architecture and the learning unit of lexical information, section 5 some experiments, and finally section 6 presents the conclusions.

2. Background

Several ontology learning approaches and systems have been proposed, which are different with each other in some dimensions, which are listed in Table 1.

Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X

Table 1. Classification of approaches to ontology learning.

Dimensions	Categories
Learning units	Word and term (single and multi-words)
Learned elements	Concepts, relationships (Taxonomy, non-hierarchical), and rules.
Data sources	Document collection, Web Dictionary and user interaction
Learning strategies	Bottom-up, top-down, and hybrid
Learning Techniques	Statistics-based, rule-base, and hybrid
Supported Knowledge	Knowledge from the web or from the users

Of these categories, the more important are the learned elements: i) Concepts: the goal is to discover new concepts as clusters of related terms. ii) Taxonomy: the goal is to discover new hyponymy relations, synonym extraction and term clustering. iii) Relations non-hierarchical: the goal is to discover new relationships between known concepts, and iv) Rules: the goal is to discover new rules, based on the discovering of behavior patterns between the words in a text (inductive learning).

With respect to learning techniques based on statistics, some of them are: Mutual information, concept mapping, and correlation analysis. With respect to learning techniques based on rules, some of them are: Heuristic patterns, dependency analysis, syntactic and semantic verb frame, concepts induction, semantically tagged corpus [4]. Some hybrid learning techniques are: parsing association, rule analysis, lexica-syntactic pattern, syntactic dependency analysis, lexica-syntactic analysis [5].

Finally, the supported knowledge refers to the source of the knowledge, from the dynamics on the web or on the user. The ontology learning approaches must have a rich knowledge representation and reasoning capabilities, and they must be able to interact with other applications present in the Web.

A variety of approaches, methods, tools and techniques have been applied to ontology learning, some of which are listed in Table 2.

Table 2. Approaches and methods to ontology learning from several sources.

Sources	Description
From text	The Aguirre and colleagues' method enriches concepts in existing Ontologies from text. They use statistical approaches to cluster topics, and the sources used for text learning is WordNet.
From dictionaries	The Hearst's method creates a thesaurus, and also enriches WordNet with new lexical syntactic relations from dictionary. The technique used is linguistic patterns and the sources used for learning is WordNet, Text.
From semi-structured data.	The Deitel and colleagues' approach enriches ontology with new concepts and relations. The main technique used is based on graph theory, and the source used for learning is RDF (Resource Description Framework) graph generated from the ontology.
From relational schema	Johannesson's method maps a relational schema to a conceptual schema. The techniques used are mappings techniques, and the sources used for learning are relational schemas.

Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X

There are tools and systems that assist the ontological engineering performing the knowledge acquisition task, which are listed in Table 3.

Table 3. Tools and systems knowledge acquisition

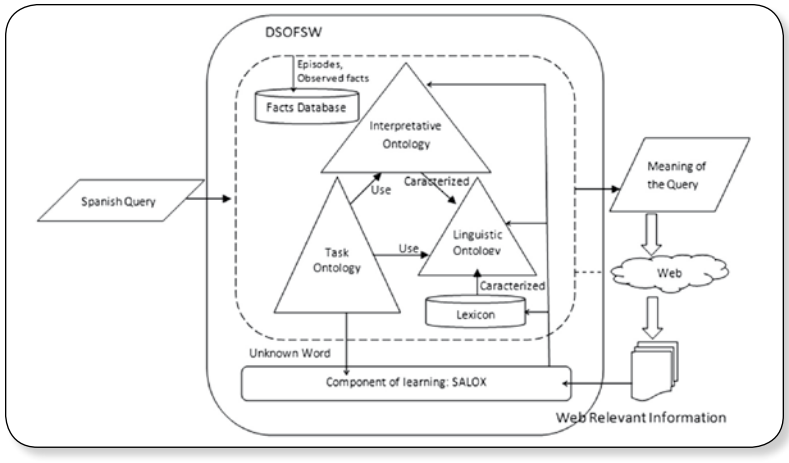
System name	Description
Doddle-OWL	(Domain Ontology Rapid Development Environment-OWL). It learns taxonomic and non-taxonomic relations using statistical methods (co-occurrence analysis), exploiting a machine readable dictionary (WordNet) and domain-specific texts.
Hasti	It is an automatic ontology building system, which builds dynamic ontologies from scratch. It learns words, concepts, relations and axioms in both incremental and non-incremental modes, starting from a small kernel (learning from scratch), using a hybrid symbolic approach: a combination of logical and linguistic information, heuristic methods, etc.
Text2ToOnto	It is an ontology learning environment based on a general architecture for discovering conceptual structures from text. They learn concepts and relations from unstructured, semi-structured, and structured data, using a multi-strategy method based on the combination of association rules, formal concepts and clustering.
WebKb	It is a system what combining statistical (Bayesian learning) and logical (FOL rule learning) methods to learn instances and instance extraction rules from world wide web documents.

The SALOX system is different from the previous approach because it integrates different learning techniques ontological and lexical in a single System; machine learning, programming inductive, fusion of ontologies, among others.

3. DSOFSW

Currently we work in the design of the Dynamic Semantics Ontological Framework for the Semantic Web (DSOFSW) [6]. DSOFSW interprets query in natural language (Spanish) to the web, and is composed by five parts (see Fig.1): a linguistic ontology for the grammar of Spanish, a lexicon for the lexical information, a database of facts about the system experiences, a task ontology for the linguistic analysis process, and an interpretative ontology of the context. The DSOFSW requires of SALOX to update the knowledge of the previous parts.

Fig. 1. Dynamic Semantic Ontological Framework



Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X

DSOFSW starts with a user query to the web; this query is semantically processed using a lexicon, a linguistic ontology, a task ontology and a domain ontology, in order to determine its meaning. The representation of the meaning of the query is converted in an OWL[1], sentence to be used by the semantic web. The information recovered of this global process is used as input to SALOX. Additionally, SALOX can receive like input lexical information during the consult interpretation when a query has an unknown word. This is the case which is studied in this paper. The interface among DSOFSW and SALOX for this case is:

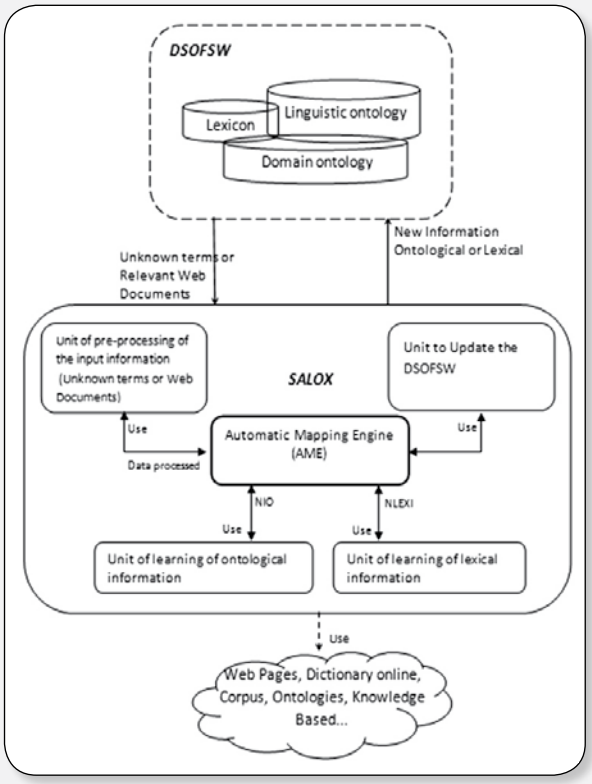
Interface: *lex_mor*[2] *(componente léxico, categoría, tipo, genero, número, modo, tiempo, aspecto, voz, persona, instancia_ontologia_linguistica*[3]*)*, where "*componente léxico*" is the place for the unknown term, the other fields are properties that are filled with information corresponding to the term.

4. Salox

SALOX is used for the adaption of DSOFSW to the dynamics of the web and of the users. SALOX considers for this two types of entries (see Fig. 2): unknown terms or relevant web document; in the first case DSOFSW invokes SALOX when it interprets the query and found an unknown terms, and the second case when DSOFSW has recovered information from the Web due to the query. From these two inputs SALOX learns: new lexical information of an unknown term for the first case, and ontological information (new concepts, relationships or properties) for the second case.

SALOX has five components: the first component is the *pre-processing unit* of the input information (unknown terms or information structured (such as databases), semi-structured (HTML or XML) as well as unstructured (e.g. textual documents)); this unit has like main task to prepare the input information for its exploitation as sources of learning. The output of this component is data pre-processed; characterized/annotated in one standard language (as XML).

Fig. 2. Learning System Infrastructure SALOX



A second component is the *unit of learning of ontological information* which contains a repository of learning techniques of ontological information; this techniques allow the learning of concepts, relations, axioms, merge ontologies, etc (for example, techniques of clustering, inductive analysis, pattern-based extraction, among others, are of interest).

A third component is the *unit of learning of lexical information* which contains a repository of learning techniques of lexical information; this component aims to find lexical

[1] Web Ontology Language: http://www.w3.org/TR/owl-features/
[2] *lex_mor* is the name defined for the defined structure for the lexicon of DSOFSW
[3] *Instancia_ontologia_linguistica* is the bridge or interface through which connect mor-phosyntactic processes and semantic analysis.

Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X

information about unknown terms (lexical-syntactic meaning). It uses online dictionaries (REA[4], Wordreference[5]), web pages, etc. In addition, it analyzes the internal structure of the unknown terms (for example, using stemming techniques (to extract sub-words, roots, affixes, etc.), among others).

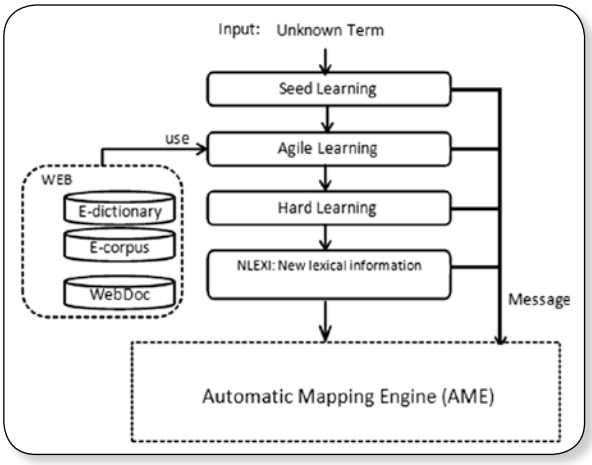
The fourth component is the unit to update the DSOFSW; this unit has like main task to upgrade the different components of DSOFSW (lexicon, its ontologies, etc.); to do this, it invokes the DSOFSW components and applies a set of specialized algorithms for updating and organizing the information in them.

Finally, the Automatic Mapping Engine (AME) component invokes the different units defined previously in order to follow the dynamics of the Web and of the users. The output of AME is new ontological or lexical information; this information is discovered by the unit of learning of ontological information (NOI: New Ontological Information) and by the unit of learning of lexical information (NLEXI: New lexical information).

4.1 Unit of learning of lexical information

This is the only unit detailed in this paper. Particularly, in this section is presented a prototype of this learning component for when there are unknown terms (see fig. 3). This learning type impacts to the lexicon. Such information (unknown terms) is characterized in the *lex_mor* function, which constitutes the interface of DSOFSW with SALOX. The goal is to learn a set of linguistic information associated with the term (e.g. linguistic categories). For example, if the word is a substantive or adjective then it learns its gender, number and type, if the word is an adverb it learns its type, and if the word is a verb it learns the mode, the tense aspect, the voice, and the person, etc.

Fig. 3. Learning unit of lexical information.



Input: *lex_mor (unknown term, null, null ...)* which is initially empty.

Output: *lex_mor (unknown term, data, data…)* filled with the new information discovered.

The unknown term is initially processed by the component of seed learning that uses stemming techniques to extract the constituent parts of the word: to extract sub-words, roots, affixes, etc. Then, both unknown term and its key parts (as the root) are sent to the agile learning component, where search and retrieval algorithms discover relevant information from the web about the different uses of the term (utilization, grammatical information, etc.). Finally, if the unknown term is a verb goes to hard learning component, to be executed by conjugation algorithms. This general procedure generates the NLEXI: new lexical information. If a given component of this unit cannot discover information then messages are generated to report that.

5. Experiments

In this case we present the behavior of our prototype for the case of the unit of learning of lexical information. Suppose that during the

[4] http://www.rae.es/rae.html
[5] http://www.wordreference.com/es/

Respuestas
Año 17
No. 2
Diciembre 2012
ISSN 0122-820X
PP: 11-19

consult interpretation there is an unknown word *"uno"*. In this case is called the *lex_mor* function:

lex_mor(uno, null, null, null, null, null, null, null, null, null, null).

The arguments *null* in the *lex_mor* is because need to be learned. The different responses from SALOX for this example are:

Fig. 4. Found Information by SALOX for this Unknown Word

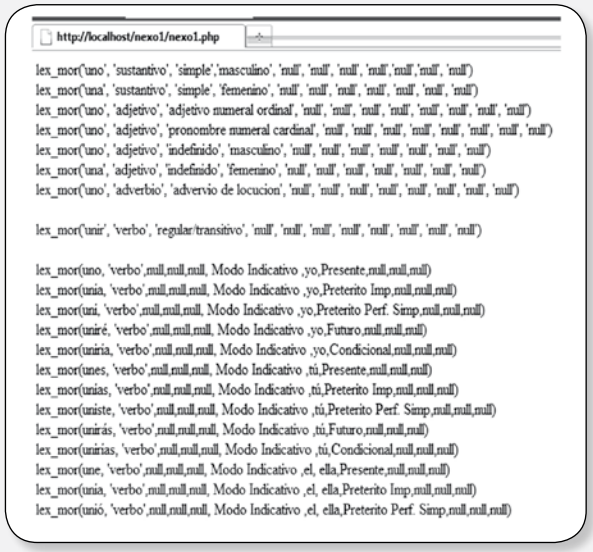


Figure 4 shows several lex_mor with different meaning possibilities for this unknown word. This information is necessary to continue with the query interpretation where the unknown word *(uno)* is included.

The first row of Figure 4 is *"lex_mor('uno', 'sustantivo', 'simple', 'masculino', 'null', 'null', 'null', 'null', 'null', 'null', 'null')"*, which means that uno is a substantive category, of type simple and genre masculine. Another possible meaning is (see line 9): *"lex_mor('uno', 'verbo', 'null', 'null', 'null', 'modo indicativo',' yo', 'presente', 'null', 'null', 'null')"*, that is uno is the first person of the present of the verb *"unir"*. We can make the same analysis for the rest of *lex_mor rows.*

Now, we evaluate the quality of SALOX, specifically to learn lexical information. At the beginning we suppose that the lexicon contains only two verbs *(ser, estar)*, some articles, adjectives and pronouns, all of which were entered manually into the lexicon. Then, we make 500 queries to prove if the lexicon grows (if we learn new terms (nouns and verbs) that were in the queries) [7]. The results are shown in Table 4.

We can see that 28% of the nouns were proper names, and normally they are not stored in traditional electronic dictionaries (WordReference.com, etc.) and 2.4% that are not learned correctly correspond to not common nouns (such as acronyms, UFPS, PDVSA, etc.). With respect to verbs, 98.8% of regular and irregular verbs were learned correctly, and only 3.2% are not learned properly (some irregular verbs that can't be described by a rule or pattern, so these verbs are not implemented in our learning component).

Table 4. Learning result.

Nouns	Learning	Not learning	Not found
Terms	174	6	70
Percent	69,6%	2,4%	28 %
Verbs	Learning	Not learning	Not found
Term	232	3	10
Percent	98,8%	3,2 %	4, %

6. Conclusions

In this paper we presented an architecture of automatic learning for DSOFSW and the design of the learning unit of lexical information. In general, DSOFSW allows the interpretation of a web query of an user in natural language. The learning component is characterized by a set of techniques of learning of lexical or ontological information. SALOX identifies unequivocally the structure to impact inside.

Our lexical learning prototype can work how a web searcher of lexical information for Spanish language. The current tools for the morphological analysis in Spanish (e.g.

STYLUS[6]. Morph syntactic tagger[7] developed by the Computational Linguistics Center at the University of Las Palmas de Gran Canarias, Morphological – Tagger[8] developed by the Language and Computing center at the University of Barcelona) have not a learning component that dynamically updates the lexicon or dictionary, our component could be used in them. Next works will be developing the rest of the components of SALOX, specifically to learn concepts, taxonomic and non-taxonomic relationships, and axioms (very important for updating the ontologies of DSOFSW) from the web documents recovered by the queries.

7. References

1. Staab, S., Studer, R.: Handbook on Ontologies, International Handbooks on Infor-mation Systems. p. 617. Berlin Springer (2005)

2. Lin, H.K., Harding J.A.: A manufacturing system engineering ontology model on the semantic web for inter-enterprise collaboration, Computers in Industry, v.58 n.5, pp.428-437 (2007)

3. Wong, W.: Learning Lightweight Ontologies from Text across Different Domains us-ing the Web as Background Knowledge. Doctor of Philosophy thesis, University of Western Australia (2009)

4. Amal, Z., Dragan G., Marek H.: Towards open ontology learning and filtering, Information Systems, v.36 n.7, pp.1064-1081. (2011)

5. Zhou, L.: Ontology learning: State-of-the-art and open issues. Information Technolo-gy and Management, 8(3), pp. 241–252. (2007)

6. Rodriguez T, Aguilar J, Puerto E.: Dynamic Semantics Ontological Framework for Web Semantics, 9th WSEAS Int.Conf. On Computational Intelligence, Man-Machine Systems and Cybernetics (CIMMACS '10), Mérida-Venezuela. pp. 91-98. (2010)

7. Rodriguez, T. Aguilar J.: Task Ontology for Lexical–Morphological Analysis of Dy-namic Semantic Ontological Framework for the Semantic Web. Conferencia Latino-americana en Informática (CLEI)'2011. Quito-Ecuador. (2011)

[6]http://stilus.daedalus.es/herramientas.php?op=pos
[7]http://www.gedlc.ulpgc.es/investigacion/desambigua morfosintactico.htm#art
[8]http://clic.fil.ub.es/demo_morfo/etiq.php?Aid=2_0_2&Aidioma=1