PROPOSAL OF A METHODOLOGICAL AND TECHNOLOGICAL DEVELOPMENT FOR AUTOMATIC ONTOLOGY EXTENSION

Jorge Cruanes, Rafael Muñoz Guillena
Department of Software and Computer Systems, University of Alicante, Apto. Correos 99, Alicante, Spain

M. Teresa Romá-Ferri
Department of Nursing, University of Alicante, Apto. Correos 99, Alicante, Spain

Keywords: Ontology extension, Ontology population, Ontology, OWL, Natural Language Processing.

Abstract: Extend an ontology is a complex task, which require a considerable amount of decision makings. This paper will study the possibility of develop an automatic ontology-driven system which will be able to extend an ontology with a satisfactory recall and precision levels, extracting information from semi-structured XML texts.

1 INTRODUCTION

Nowadays, most of the information is textual Nowadays, most of the information is textual information is textual (news, medical brochures, reports, legal texts, etc.). There is a social need to have the information required in the shortest time, but this information, even it is digital, is thought to be processed by humans and little or none prepared for computer processing.

Ontologies are one of the given solutions to represent information in meaningful structures for machines. An ontology consists in a serial of concepts and their relationships, which can represent the knowledge domain. However, the current ontologies do not cover all needs of the knowledge representation. That is why its extension is needed.

The extension of ontology has become a major bottleneck (Mehrnoush et al., 2004; Adrian et al., 2009), and it is extremely complicated, but necessary to maintain a constant updating of an ontology, both general concepts (known as classes or nodes) and specific terms (known as instances or concepts of a specific nature).

This paper is divided in six sections, beginning with this introduction, which ends with a description of the objectives and how we expect to accomplish them. The section 2 provides a summary of the state of the art. In the section 3 it is described the proposed system. The 4th section shows the future work to follow on this research line. Section 5 contains the conclusions reached during the study and finally, in section 6, there are the references.

1.1 Objectives

This paper is a proposal of methodological and technological development, capable of extending an ontology automatically.

The overall objective is, from a collection of semi-structured[1] documents in XML format, to extract the necessary structured information to extend the reference ontol-ogy (it is also known as base ontology). All these processes are performed automati-cally and applying Natural Language Processing (NLP) techniques and complemented with external linguistic resources (dictionaries, lexicons, thesaurus, etc.).

Other goals that define the research line that begins with this work are:

[1] Semi-structured documents are to those, which although they have a certain hierarchical structure and labels, there is a lack of semantic information (both formal and explicit relations between concepts).

Cruanes J., Muñoz Guillena R. and Romá-Ferri M..
PROPOSAL OF A METHODOLOGICAL AND TECHNOLOGICAL DEVELOPMENT FOR AUTOMATIC ONTOLOGY EXTENSION.
DOI: 10.5220/0003117304710475
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 471-475
ISBN: 978-989-8425-29-4

- Distinguish if the extracted information should be considered as concepts, instances or relations by using the reference ontology, and incorporate them into the ontology.
- Determine recall and precision of the identification and extension of the reference ontology.
- Create a modular, multi-platform and distributed system.
- Determine the efficiency of the system.

Once the objectives and perspectives of the system have been established, next section will provide the current status of the issue and its shortcomings.

2 STATE OF THE ART

2.1 Ontology Definition

One of the most referenced definitions of ontology in the literature is the one given by Gruber (1993): "An ontology is a specification of a conceptualization". Several authors qualified this definition since then, including Studer and team (1998), who stated that "[...] conceptualization refers to an abstract model of some phenomenon in the world to be identified by relevant concepts of that phenomenon". We could say, therefore, that an ontology is an explicit representation of the ideas of the real world, where these ideas are represented formally by their characteristics and relationships between them.

In an ontology, the knowledge can be represented by concepts, relations and in-stances. Concepts represent generic information in the real world, establishing their characteristics and properties. Instances are specifications of concepts, and have their characteristics. The relationships provide information on how concepts relate to each other and, therefore, how instances can be related.

2.2 Ontology Extension

The extension of an ontology can be done in horizontal or vertical way. It is said that an ontology is extended horizontally when we add generic information as ontology concepts. On the other hand, we say that an ontology is extended vertically when we add specific information (instances or terms). This type of information is known as leaves or instances of the ontology. This vertical extension of the ontology, which adds specific knowledge, is no longer part of the ontology itself, but constitutes the knowledge base ontology.

It is important to understand that ontology is where the knowledge represented by nodes and their relationships is explicit. The base of knowledge requires the existence of an ontology in order to understand the terms it represented. Chandrasekaran and team (1999) define this distinction telling that the ontology is not the instances them-selves, but the conception of the concepts that are trying to capture.

To illustrate the extension of an ontology, here is an example of extension, starting from the base ontology illustrated in the figure 1.

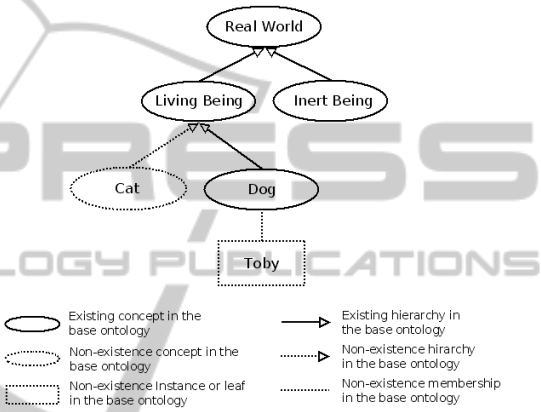

Figure 1: Example of an ontology with four concepts. There is also one concept and one instance that will be added to the ontology, which are dotted.

The ontology used as base ontology with the concepts of 'Living being' be 'Inert being' and 'dog', which is dependent on 'Living Being'. All of these terms inherit from the node 'Real World', which is the root node of the ontology. This dependence is given by the relationship 'is a', represented by a solid line arrow with the white arrowhead. This relationship allows the child concept (the node at the origin of the arrow) to inherit the properties of the parent (the node to whom the inheritance arrow points).

Human knowledge is constantly evolving and constantly finding new concepts (such as the HLC or the iPad), and at some point we will realize that the ontology is not enough to represent all the knowledge required, so the ontology must be updated. Continuing with our example, it was decided to extend the ontology with the ideas of 'cat' and 'Toby'. In the first case ('cat') is a concept, while the second ('Toby') is the name of a specific dog.

Making the extension we decide that 'cat' will become a node in the ontology, because it represents

a generic concept. The right place will be dependent on 'Living being', inheriting its features, just as does the concept of 'dog'. In this situation, as explained above, 'cat' will inherit the properties and characteristics apply to the concept 'Living being'.

Along with the enlargement process, 'Toby' will be included as an instance. Therefore, the ontology is extended adding 'Toby' to this specificity level of the concept 'dog'. The instance 'Toby' has the same characteristics as 'dog'. If the node 'dog' had an attribute called 'hair colour' and it accepts a string as a value, 'Toby' will be characterized by the value of 'brown' for that attribute.

2.3 Related Work

For the extension of an ontology or a knowledge base there are various systems and projects in development.

Regarding the techniques used in information extraction, there are many similari-ties with those applied in extension. The nowadays information retrieval systems use Part-Of-Speech (POS) tagging tools (Hamdi et al., 2008) and pattern systems (Sun et al., 2007, Thiam et at., 2008 and 2009, Simon Cuevas et al., 2009) for entities recog-nition or filtering data, while the identification of synonyms, abbreviations and acro-nyms used additional resources such as dictionaries and WordNet[2] (Makki et al., 2008; Simon-Cuevas et al., 2009). In systems that extract information from databases (Roma-Ferri, 2009b), for the entities recognition or data filtering they use ad-hoc extraction systems.

Once the information has been extracted, the systems store it in an ontology or knowledge base (Makki et al., 2008, Thiam et al., 2008; Simon-Cuevas et al., 2009). To achieve the storage, it is needed to match the information that should be repre-sented on the knowledge base or ontology, and the one extracted from the source texts, trying to get where to add that new information to make the extension. There are some different NLP techniques, such as mapping by comparing labels or values (Sun et al., 2007, Hamdi et al., 2008), based on rules (Simon-Cuevas et al., 2009) or using external tools (Deléger et al., 2006).

Almost all systems are automatic, except for some used in fields such as medicine, due to the sensitivity of information processed, as the system created by Sun and team (2007). This semi-automation requires an expert to validate and filter the matches made by the system before being added to the knowledge base.

The work of Makki and team (2008) is one of the most interesting according the frame of this paper. They provide a system to populate an ontology in the risk man-agement domain. Their system works with unstructured documents (technical specifi-cations of Chemistry) and uses NLP tools in the extraction (POS tagger and the use of WordNet for the words expansion found in the text). The words expansion is a tech-nique whereby are obtained semantically similar terms to those used as base, increas-ing the chances of finding semantic matches.

Works oriented to ontology extension (in literature, Ontology Learning) are achieving good recall and precision levels. However, to achieve those levels, some systems require an important expert supervision (Sun et al., 2007). Also, another issue that affects recall and precision are the dependence to particular language or belong-ing to a specific domain (Makki et al., 2008; Simon-Cuevas et al., 2009).

Therefore, there remains the need of a fully automatic system, context and lan-guage independent, capable of extending an ontology base.

3 SYSTEM DESCRIPTION

Having identified the shortcomings, we propose a system capable of extending an ontology, automatically, from an ontology provided as reference and a collection of semi-structured texts. The system will work with semi-structured documents in XML format, will be independent of domain ontology, and even the language used.

The system will work in phases from semi-structured documents in XML format, (i) extracting information driven by a reference ontology (the base ontology), (ii) processing it to being structured and adding semantic information and, (iii) finally, extending the base ontology.

3.1 System Functionalities

The system is split in three main modules: the Textual Information Extractor (EIT), the Engine and the Ontology Generator (as shows figure 2).

[2] An on-line browser can be accessed from the next URL: http://wordnetweb.princeton.edu/perl/webwn.

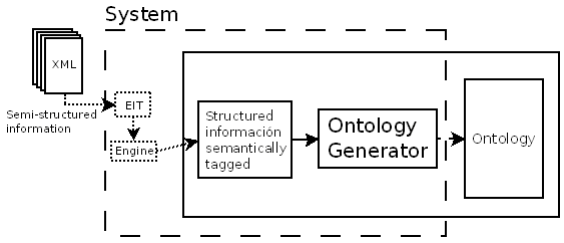

Figure 2: Internal architecture of the system and the information flow from semi-structured XML text documents to the ending output in OWL/RDF, which is solid-framed the part of the system focused of this paper.

- **Textual Information Extractor (EIT).** This is the module responsible for extracting the relevant information from semi-structured documents provided as input. A POS tagger will deal with the text lo-cated in the input documents, so we will have syntactic and lexical information that allows us to identify tokens candidates to become concepts, terms or relationships.

Using NLP techniques, the existing concepts in the base ontology will be expanded in order to be sought in the text. The expansion of the terms will be made through the use of the existing synonyms in the base ontology and those founded in diction-aries and WordNet. If the system founds polysemy it will be showed up, returning all the groups to which they may belong.

- **Engine.** It collects the EIT output to decide, depending on the reference ontology, what in-formation will be slated to become instances, concepts or relationships, and which of them will be accepted or discarded. In the case of polysemy, this module will de-cide the proper meaning based on the context of the token.

As output, the system will provide the semantic information needed to include the terms, concepts and relationships in the ontology.

- **Ontology Generator.** This module is responsible for collecting the Engine's classification to extend and populate the ontology with the appropriate format, generating the final output of the system. Thanks to the Engine's output this module will decide where to place new instances and new concepts, and establishing the appropriate relationships.

The enlargement process of the ontology consists of the inclusion of the concepts, terms and relationships based on information supplied by the Engine, and a subse-quent post-processing.

The concepts will be introduced where they belong thanks to the semantic informa-tion provided by the Engine, keeping the tree structure given by the OWL/RDF schema. Thus, the concept 'cat' according to figure 1, this module will create a node placed as child of node 'Living being' and sibling of 'dog'.

The Ontology Generator post-process will verify the absence of loops in the inheri-tance relationships. If there are, all classes in the loop will be declared as equiva-lents. To perform this check it will use the algorithm of Tarjan (1972), widely used to find cycles in directed graphs.

This module will also check if there is any relationship between two instances but it is not between the classes to which they belong. In this case, it will be added.

The Ontology Generator's output is split into two files: one for the ontology itself and the other one for its instances. This is to facilitate the maintenance of the ontol-ogy by experts, other automated systems or software tools. This design will also im-prove efficiency in the management of ontology by software tools, since the inclu-sion of the instances in the same file will generate very big documents, such as On-toFIS ontology (Roma-Ferri, 2009a).

4 FUTURE WORK

The first objective of continuity of this project is to implement the system designed, test it and evaluate it. The implementation will be made in a high-level language (C #, Java, etc.) and there will be used web services as application interface. The use of web services will achieve the goals of modularity and make a distributed and plat-form-independent application.

Once developed, the system will be tested on the extension and population of a base ontology. In order to do this, we will use a collection of controlled data, so we will have complete control about: (i) how accurate data are entered into the system, (ii) what data are accepted or discarded on the reference ontology, and (iii) what we expect as a output ontology.

To check the extension process of the reference ontology (both the horizontal and vertical), we have established a systematic evaluation. This routine will include a pre-test and post-test of the reference ontology from a control question bank, by compe-tency questions (Gruninger, 1995; Roma-Ferri, 2009b). The system should achieve values of

coverage and accuracy in detection of concepts, terms and relationships in semi-structured texts over a 75% before working with unstructured texts.

5 CONCLUSIONS

This paper proposes a methodological and technological development capable of extending an ontology automatically. To solve this problem we have presented a system that uses NLP techniques and tools (such as POS taggers, textual information extraction, or token matching), which have been proved in previous projects.

This proposal also indicate a systematic tests and milestones to demonstrate the quality of the methodology and technology presented.

ACKNOWLEDGEMENTS

This paper has been supported partially by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educació - Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/286).

REFERENCES

Adrian, B., Hees, J., Van Elst, L., Dengel, A.: iDocument: Using Ontologies for Extracting and Annotating Information from Unstructured Text (2009)

Chandrasekaran, B., Johnson, T. R., Benjamins, V. R.: What are ontologies, and why do we need them?. *IEEE Intelligent Systems*, 14 (1) (pp. 20-26) (1999)

Deléger, L., Merkel, M., Zweigenbaum, P.: Using Word Alignment to Extend Multilingual Medical Terminologies (2006)

Gruber, T. R.: Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL-93-04, *Stanford Knowledge Systems Laboratory Report* (1993)

Gruninger, M. y Fox, M. S.: Methodology for the Design and Evaluation of Ontologies. In: *Proceedings of the Workshop on Basic Ontological Issues Expoert Systems*. Liebowitz, ed. CDR Press (1995)

Hamdi, F., Zargayouna, H., Safar, B., Reynaud, C.: TaxoMap in the OAEI 2008 alignment contest. Third International Workshop On Ontology Matching (OM2008) (2008)

Makki J., Anne-Marie, A., Prince V.: Ontology Population via NLP Techniques in Risk Management. World Academy of Science, Engineering and Technology 43 (2008)

Mehrnoush, S. y Barforoush, A. A.: The State of the Art in Ontology Learning: A Framework for Compariso. The Knowledge Engineering Review 18 (pp. 293-316) (2004)

Romá-Ferri, M. T.: OntoFIS: Tecnología ontológica en el dominio farmacoterapéutico. Tesis doctoral. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante, Alicante, Spain (2009a)

Romá-Ferri, M. T.; Cruanes, J.; Palomar, M.: Quality indicators of the 'OntoFIS' pharmacotherapeutic ontology for semantic interoperability. IADIS International Conference e-Health 2009. In: *Proceedings of the IADIS International Conference E-Health 2009*. ISBN: 978-972-8924-81-2. Ed. IADIS. (pp. 107-114) (2009b)

Simón-Cuevas, A. J., Ceccaroni, L., Rosete-Suárez, A., Suárez-Rodríguez. A.: A Formal Modeling Method applied to Environmental-Knowledge Engineering. International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009), *IEEE Computer Society Press* (pp. 1132-1137) (2009)

Studer R., Benjamins, V. R., Fensel, D.: Knowledge Engineering: Principles and Methods. *IEEE Transactions on Data and Knowledge Engineering* 25 (1-2) (pp. 161-197) (1998)

Sun, J., Bai, X., Li, Z., Che, H., Liu, H.: Towards a Wrapper-Driven Ontology-Based Framework for Knowledge Extraction. *Lecture Notes in Computer Science* N. 4798, (pp. 230-242) (2007)

Tarjan R.: Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing* 2 (pp. 146-160) (1972)

Thiam, M., Pernelle, N., Bennacer, N.: Contextual and Metadata-based Approach for the Semantic Annotation of Heterogeneous Documents (2008)